Universidad
Carlos III de Madrid

Clasificación de patrones de siniestros aplicando técnicas de agrupación y segmentación

Gabriel Anca Corral

26 de Septiembre de 2015

Agradecimientos

Es imposible expresar con palabras el sentimiento de gratitud que siento hacia todas aquellas personas que me han prestado apoyo moral e intelectual durante el transcurso de este maravilloso periodo de mi vida, que comenzó el lunes 7 de septiembre de 2009. Este proyecto no es más la muestra de que todo el esfuerzo dedicado durante más de seis años de carrera ha merecido la pena.

Me gustaría agradecer en especial a María Diéguez por todo el apoyo que me ha prestado a lo largo de todo este periodo, demostrando estar ahí para apoyarme, ayudarme a ver la parte positiva de todo y conseguir que siga adelante en todos los momentos difíciles. Gracias por estar ahí siempre.

No se puede olvidar agradecer a mis padres, Pedro Anca y María Teresa Corral, y a mi hermano, Sergio Anca, por darme la oportunidad de emprender esta aventura y por todo el ánimo y apoyo que me han dado estos años. Sin vosotros no habría sido posible.

Ha sido esencial la colaboración en este proyecto de mis tutores, Irene Albarrán y Miguel Ángel Patricio, los que desde el primer momento se han volcado en guiarme para conseguir que este proyecto fuera un éxito. La orientación proporcionada por Antonio Berlanga en todo momento ha sido, sin duda, otro pilar fundamental de este proyecto. Os estoy muy agradecido.

Doy las gracias a mis amigas y compañeras Nuria Villa, Alexandra Escobedo, Laura Acosta y Esther Matesanz por todos estos años. Hemos estado juntos en los buenos y en los malos momentos, hemos trabajado en equipo, nos hemos ayudado los unos a los otros y, sobre todo, hemos conseguido que el camino haya sido más fácil y más ameno.

Gracias a mis amigos más cercanos Alejandro, Claudia, David, Fernando, Javier, José María, Mario, Pablo, Rebeca y Rubén, por todo el apoyo y ánimo que me habéis transmitido durante todos estos años.

Agradezco también al Campus de Colmenarejo de la Universidad Carlos III de Madrid por hacer todo lo posible para que nos desarrollemos como estudiantes y como personas, y poner a mi disposición todo lo que necesito para ello.

Es imposible agradecer en una sola página a todas las personas que han colaborado para que esto sea posible, pero sin duda todas las personas que han pasado por mi vida durante estos seis años han sido muy importantes para mí.

Tabla de contenidos

Índice de ilustraciones

Índice de tablas

Índice de ecuaciones

Índice de fragmentos de código

1. Introducción

1.1. Contexto

Los vehículos a motor son un elemento indispensable para el día a día de la sociedad actual, permitiendo el transporte de personas y mercancías. De acuerdo con el último Anuario de la Dirección General de Tráfico [1], al final de 2014 había casi 31 millones de automóviles registrados en España. A pesar de los grandes avances en infraestructura y en seguridad tanto en los vehículos como en las carreteras, todavía se produce un elevado número de accidentes, que causan daños tanto personales como materiales. El riesgo de que estos accidentes ocurran y el coste incierto de los mismos motivan la existencia de los seguros de automóviles.

El seguro de automóviles cubre los riesgos derivados al uso y circulación de los mismos. Principalmente incluye coberturas relativas a la responsabilidad civil del asegurado por daños que se causen a terceras personas o a sus bienes, aunque también puede incluir cobertura a los daños al conductor, a los pasajeros del vehículo o al propio vehículo asegurado.

De esta forma, el asegurado, en caso de que ocurra un siniestro cubierto por la póliza del seguro, transfiere la obligación generada por la responsabilidad civil del asegurado a la aseguradora, siendo ésta última la que corre con los gastos derivados de sus siniestros. La prima pagada por los asegurados es utilizada para costear los siniestros, lo que hace crucial el cálculo de la misma para evitar riesgos de insolvencia en las aseguradoras.

El cálculo de la prima de un seguro se basa en múltiples parámetros muy complejos, unos se pueden estimar previamente a la firma del contrato: factores relativos al vehículo, al conductor y a la conducción (*a priori*); y otros posteriormente a la firma del contrato: historial de siniestralidad, multas, etcétera (*a posteriori*).

La normativa más reciente a nivel europeo creada por la Comisión Europea (conocida comúnmente como normativa sobre Solvencia II) está formada por la *Directiva 2009/138/CE del Parlamento Europeo y del Consejo, de 25 de noviembre de 2009, sobre el seguro de vida, el acceso a la actividad del seguro y de reaseguro y su ejercicio (Solvencia II)* y el *Reglamento Delegado (UE) 2015/35 de la Comisión, de 10 de octubre de 2014*. Esta directiva define unos requisitos de capital en función del riesgo asumido por las entidades aseguradoras, de manera que se reduzca el riesgo de insolvencia. Dada la inminente entrada en vigor de esta norma, es crucial mejorar los métodos de cálculo de las primas a cobrar a los asegurados, para poder cumplir estos requisitos de capital repartiendo las primas de una forma adecuada y justa entre cada uno de los asegurados de las compañías.

Es de suponer que cada aseguradora debe manejar una base de datos de peritajes de siniestros, con los que pueda llevar un registro interno de los costes y las circunstancias de cada una de las reparaciones a vehículos que hayan realizado a lo largo del tiempo. Una de las compañías aseguradoras más importantes de España en la rama de automóviles, la Mutua Madrileña, en su Informe Anual de 2014 [2], y gestionó un total de 1.304.000 partes de accidente en la rama "Auto", teniendo una cuota de mercado es del 12,5%. A pesar de que muy pocas compañías hacen pública esta información, podemos comprender que la cantidad de información generada por las compañías aseguradoras cada año crece a pasos agigantados. En un escenario en el que se disponga de estos datos a lo largo del máximo periodo de tiempo posible por todas estas compañías aseguradoras, podríamos sacar conclusiones tremendamente útiles para las compañías de seguro. Un buen análisis de esta información podría ayudar a las aseguradoras a, entre otras cosas, mejorar el cálculo de las primas -concretamente la parte del cálculo de la prima basado en las características del vehículo-, detectar el fraude y sacar conclusiones estadísticas muy interesantes.

Estos datos tendrían una dimensión tan grande que resultaría inmanejable para cualquier ser humano e incluso para computadores domésticos. Se necesitan herramientas avanzadas de análisis de datos para poder extraer conclusiones en un tiempo razonable. Afortunadamente, en la era que vivimos, denominada "Era de la Información", existen numerosas herramientas especializadas para analizar y clasificar la enorme cantidad de información que se genera en todo el mundo cada segundo.

A pesar de la existencia de todas estas herramientas, éstas están en continuo desarrollo, y su uso todavía no está muy extendido y está escasamente documentado. Trabajar con este tipo de sistemas no es una tarea fácil, y requiere mucha investigación y conocimientos aplicados para conseguir resultados válidos.

1.2. Objetivos

Una importante empresa de peritaje de seguros nos ha cedido una base de datos que contiene información detallada sobre más de 390.000 partes de accidente que corresponden a un modelo concreto de vehículo (delimitado por marca, modelo y año de fabricación). Estos datos contienen información sobre cada una de las piezas que han tenido que ser sustituidas, reparadas y/o pintadas, así como el coste de reparación desglosado por coste de piezas nuevas, coste de mano de obra y coste de pintura.

Este proyecto se basará en el análisis de una base de datos de partes de accidente para conseguir agrupar los siniestros de cada marca y modelo de vehículo según dos criterios distintos:

- Según la severidad del siniestro: se crearán grupos basados en el coste económico de la reparación del siniestro. Dentro de cada uno de estos grupos, se volverá a realizar una separación en grupos, esta vez buscando la existencia de zonas de impacto predominantes en cada uno de los grupos de severidad.
- Según la zona de impacto: se crearán grupos basados en las zonas que hayan resultado dañadas en cada siniestro. A partir de la obtención de dichas zonas, se crearán distintos grupos de severidad para cada una de ellas, analizando qué niveles de coste puede tener cada uno de los grupos.

Se crearán también árboles de decisión que nos permitan, dado un siniestro nuevo, decidir a qué grupo corresponde sin tener que ejecutar el algoritmo de agrupación de nuevo, así como detectar fraude en el peritaje.

Esta información tiene múltiples aplicaciones: por un lado conseguiremos mejorar el cálculo de la prima de los seguros de automóvil. En concreto se tratará de mejorar una parte de la prima derivada de la segmentación de riesgos o *a priori*: aquella derivada de las características del vehículo. Esto se llevará a cabo basándonos en los grupos de severidad y zonas de impacto para cada marca, modelo y año de fabricación del vehículo, por otra parte se mejorará la detección de fraude en la peritación de siniestros, con ayuda de los árboles de decisión, dado que un parte de siniestro que no encaje con algún grupo ya existente puede querer decir que haya habido algún tipo de fraude, y deberá ser revisado en detalle.

Adicionalmente, con la ayuda de los árboles de decisión se conseguirá retroalimentar esta base de datos, pudiendo fácilmente añadir todos los nuevos siniestros a esta base de datos consiguiendo que ella misma se nutra con la nueva información.

Dada la dimensión de los datos de que disponemos, para todos estos procesos se utilizarán algoritmos de clustering, que serán aplicados de manera distribuida entre varios computadores, buscando conocer el funcionamiento de este tipo de software y optimizar el proceso, de manera que el agrupamiento de los datos tome el menor tiempo posible y, además, se puedan realizar el mayor número de análisis sobre los mismos para extraer mayores y mejores conclusiones. También será necesario evaluar los distintos algoritmos de agrupamiento de datos, y de creación de árboles de decisión, así como las distintas implementaciones que haya de los mismos, para utilizar aquel que presente mayor eficiencia.

Debido al tipo de procesos que se van a realizar, este estudio también tendrá como objetivo evaluar el rendimiento de las tecnologías de minería de datos ejecutadas sobre sistemas de

computación distribuida. Esto no sólo permitirá conocer el rendimiento de este tipo de software en conjunto con los citados algoritmos, sino que también ayudará a comprender de qué manera influye al rendimiento la adición o sustracción de computadores de distintas características al clúster, en función del tipo de datos que se estén analizando. De esta forma, esto no sólo será aplicable en el estudio que se está realizando sino también podrá ayudar a tomar decisiones en otras materias.

Aunque la información de que disponemos es sólo de una marca y modelo de vehículo en concreto, se pondrá especial ímpetu en conseguir que la infraestructura y el software empleado en este proyecto sea escalable y válido para nuevas bases de datos de otras marcas y modelos, así como para nuevos siniestros que vayan aconteciendo en el futuro.

1.3. Marco regulador

Debido a que los datos en los que se basa este estudio han sido cedidos por una entidad privada, y que los datos que contiene en su interior son de personas y empresas particulares, de acuerdo a la Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal, se ha prestado especial atención para borrar todos los rastros de datos personales que existan en dicha base de datos, asegurándonos de que cada una de las entradas de la misma son completamente anónimas.

Por otra parte, y de acuerdo a la misma ley y con los requisitos específicos de la empresa que ha cedido los datos, debido a la firma de un contrato de confidencialidad, no se incluirá en este estudio ninguna información personal sobre la empresa que ha cedido los datos ni sobre los mismos datos.

2. Situación actual

Es conveniente, antes de entrar en materia, conocer ciertos aspectos del entorno que rodea este estudio, de manera que se pueda entender a la perfección todo su contenido. En primer lugar, se darán unas nociones básicas sobre los seguros y en concreto el seguro de automóvil. A continuación se describirá el proceso de tarificación así como la gestión de siniestros y el fraude en este tipo de seguros. Más adelante, se tratarán aspectos relacionados con el análisis de datos, como los que se utilizarán en este estudio, abordando la minería de datos, el aprendizaje automático y los algoritmos de agrupamiento o clústering.

2.1. Introducción al seguro de automóviles

2.1.1. Introducción a los seguros

Pérez Torres [3] define el seguro como ≪la institución en que las personas que están expuestas a un riesgo agrupan sus recursos en un fondo común para hacer frente a las consecuencias económicas negativas que se producirían para aquellas en las que el hecho constitutivo del riesgo ocurra realmente≫. De una forma más sencilla se podría decir que una aseguradora es la que se hace cargo de los costes derivados de los siniestros que ocurran a los asegurados, utilizando para ello la cuota que se ha cobrado a priori a todos los asegurados.

Existen diversos tipos de seguros, dependiendo de aquello que se esté asegurando:

- Seguros de personas: cubren riesgos que afectan a la vida, salud o integridad de una persona o grupo de personas.
- Seguros de daños o sobre las cosas: cubren aquellos riesgos que se produzcan en los bienes del asegurado.
- Seguros de patrimonio o de responsabilidad: cubren los riesgos que produzcan obligaciones patrimoniales a un asegurado.
- Seguros multirriesgos: incluyen varias coberturas de entre las anteriores.

2.1.2. El seguro de automóvil

El seguro de automóvil o seguro de vehículos a motor es uno de los seguros más influyentes en la vida económica y social de los países desarrollados. Los vehículos a motor forman parte de la actividad diaria de la mayoría de nosotros. Sin embargo, estos vehículos generan una serie de riesgos, tanto personales (sobre los ocupantes de los mismos) como materiales (sobre el propio vehículo y los bienes que en él se transportan), que dan lugar a la existencia del seguro de automóviles.

Este tipo de seguro surge con el fin de afrontar el coste de reparación de los daños producidos por el uso de vehículos de motor por el uso y circulación de los mismos. En los vehículos a motor

se engloban, por ejemplo, turismos, furgonetas, ciclomotores, motocicletas, autocares, camiones y tractores. En cuanto al uso y circulación de los vehículos, se incluyen los hechos ocurridos en cualquier vía o terreno, ya sea público o privado, siempre y cuando sean aptos para la circulación de vehículos de motor. Dado que según el Código Civil es el conductor causante de los daños el que se deberá hacer cargo de los mismos, los conductores suscriben un seguro de responsabilidad civil para transferir el daño que estos siniestros pueden producir sobre el patrimonio del conductor. De esta forma, el conductor está transfiriendo dicho riesgo a un seguro de automóvil.

La norma que regula en este momento los seguros de automóvil es la Ley sobre responsabilidad civil y seguro en la circulación de vehículos a motor [4], la cual surgió en 1962, se comenzó a aplicar en 1968 y ha sido revisada en numerosas ocasiones, siendo la última el 1 de enero de 2014. Uno de los artículos más importantes de esta ley es el segundo, que establece la obligatoriedad de la suscripción de un seguro de responsabilidad civil para todos los vehículos con estacionamiento habitual en España.

La responsabilidad civil recientemente mencionada se define como la obligación que tiene una persona de reparar los daños y perjuicios ocasionados a la integridad física o los bienes de un tercero ya sea por acción u omisión. La reparación de estos daños será realizada por la persona culpable de este siniestro.

El seguro de responsabilidad civil se enmarca dentro de los seguros patrimoniales, es decir, cubre el riesgo de que le sea reclamado responder con su patrimonio las personas dañadas por sus acciones u omisiones, respondiendo la aseguradora en lugar del asegurado del pago de las obligaciones generadas por el asegurado. Aun así, los seguros de automóvil habitualmente incluyen coberturas para otro tipo de riesgos (vida, salud o integridad de los pasajeros, daños ocasionados en el propio vehículo, asistencia en viaje, protección jurídica, etc.), por lo que habitualmente se trata de seguros multirriesgos.

Este seguro se diferencia de los seguros de responsabilidad civil convencionales en tres aspectos:

- Principio de responsabilidad objetiva. Este principio permite que exista responsabilidad civil sin que exista culpa o negligencia, a diferencia de la responsabilidad civil subjetiva. Dicho de otra manera, la responsabilidad objetiva se basa en la causalidad frente a la responsabilidad subjetiva, la cual se basa en la culpabilidad. De esta forma se consigue que el resarcimiento de los daños y perjuicios causados sea más rápido al no tener que determinar la culpabilidad o negligencia de los implicados en un siniestro.
- Carácter obligatorio del seguro de automóviles. Por lo general la suscripción de un seguro de cualquier tipo es voluntario, mientras que en todos los países de la Unión Europea y

por ende en España, mantener un seguro de responsabilidad civil para cada automóvil es de carácter obligatorio.

- Determinación objetiva de los daños personales. La forma de evaluar los daños producidos difiere entre los seguros de automóvil y el resto de seguros en que, a pesar de que la indemnización será objetiva y dictada por un juez en ambos, en el seguro de automóvil se utilizan reglas objetivas para valorar los daños personales. En España existe un Baremo de valoración para los daños personales del que hablaremos con más detenimiento cuando tratemos la peritación de siniestros en el apartado 2.3.

A pesar de que hemos visto que este seguro puede tener varias coberturas, el fin principal del mismo es cubrir la responsabilidad civil del asegurado generada por el uso y circulación de los vehículos de motor. Con respecto a esta responsabilidad, existen dos tipos: la obligatoria y la voluntaria. La responsabilidad obligatoria se destina a la cobertura de daños personales o materiales causados a terceras personas, dentro de un límite establecido por la ley. Sin embargo, la responsabilidad voluntaria amplía dichos límites y ofrece otras coberturas.

Por tanto, las coberturas que habitualmente puede ofrecer un seguro de automóvil son las siguientes:

Responsabilidad civil obligatoria.

En España, la Ley sobre responsabilidad civil y seguro en la circulación de vehículos a motor regula el seguro obligatorio de automóviles, el cual se encarga de cubrir la responsabilidad civil del propietario o conductor del vehículo por las lesiones corporales o daños materiales que se puedan causar a terceros.

Actualmente esta cobertura se limita a 350.000 euros de daños corporales por víctima, 100.000 euros por daños materiales, gastos completos de asistencia médica, farmacéutica y hospitalaria y gastos de entierro y funeral en caso de muerte. Sin embargo, la legislación excluye de esta cobertura los daños y perjuicios causados al causante del siniestro, ni a su vehículo ni a los bienes transportados en él.

Existen otras dos excepciones con respecto a las coberturas de este seguro: si el conductor circulaba bajo los efectos del alcohol o de otras sustancias estupefacientes, o si el conductor circulaba con un vehículo robado. En el primer caso, la aseguradora adelantará el dinero, y luego tendrá derecho a reclamarlo al asegurado. En el segundo caso, será el Consorcio de Compensación de Seguros el encargado de indemnizar a los damnificados.

Responsabilidad civil voluntaria.

Muchas aseguradoras ofrecen la posibilidad de aumentar los límites establecidos por el SOA, a través del seguro voluntario de automóviles (SVA). Estos nuevos límites serán acordados por la aseguradora y el asegurado y vendrán establecidos en el contrato.

Al igual que en el seguro obligatorio, existen ciertos riesgos excluidos del SVA, como por ejemplo las multas, la responsabilidad civil contractual y la responsabilidad por los objetos transportados por el vehículo si el transporte de los mismos no se adecúa al Código de circulación. Sin embargo, estas coberturas pueden ser incluidas expresamente en el contrato.

Protección jurídica.

El seguro de protección jurídica suele incluir la defensa del asegurado o del conductor del vehículo mediante abogados o procuradores, así como los gastos judiciales y el depósito de fianzas.

Por otra parte, esta cobertura suele incluir la reclamación de daños personales o materiales que un tercero haya causado al asegurado.

Asistencia en viaje.

El seguro de asistencia en viaje incluye como beneficiarios al tomador de la póliza, las personas de su unidad familiar, así como el conductor y los pasajeros del vehículo. Las garantías incluyen gastos médicos, transporte, gastos de los acompañantes, desplazamiento de familiares, repatriación, etc.

Daños propios al vehículo.

Esta cobertura es la primera de todas las mencionadas que no cubre el patrimonio del asegurado. En este caso lo que se cubre son los daños en el vehículo propio por vuelco, caída, choque, incendio, explosión, rayo, robo y/o rotura de cristales. Cada entidad aseguradora ofrece estas garantías, según el tipo de seguro contratado, de forma conjunta o por separado. Habitualmente en los seguros a todo riesgo, se incluyen todas estas coberturas en conjunto.

Riesgos personales.

Este tipo de seguro cubre los riesgos personales derivados de los accidentes ocasionados por el uso y circulación de los vehículos. Este seguro incluye indemnizaciones limitadas por el contrato del mismo para accidentes en los que se produzcan lesiones corporales, incapacidad o muerte del conductor o los pasajeros del vehículo asegurado. Este seguro

suple la falta de cobertura de los daños personales del culpable del accidente por parte del seguro de responsabilidad civil.

Aparte de las coberturas mencionadas con anterioridad, dependiendo del contrato de seguro existen otras coberturas complementarias al seguro de automóviles que son incluidas en algunas ocasiones, como por ejemplo la cobertura por retirada del permiso de conducción, cobertura de equipajes o la cobertura de limpieza y acondicionamiento del vehículo en caso de daños producidos por transporte de víctimas heridas de un accidente de circulación.

Por lo general, las compañías aseguradoras ofrecen tres modalidades distintas de seguros, combinando las coberturas citadas anteriormente:

- A terceros: incluye la responsabilidad civil obligatoria, responsabilidad civil voluntaria, protección jurídica, asistencia en viaje y riesgos personales.
- A terceros ampliado: incluye todas las coberturas de un seguro a terceros y daños propios por rotura de cristales, incendio y/o robo.
- A todo riesgo: incluye las coberturas de un seguro a terceros y todos los daños propios del vehículo (en ocasiones aplicando una franquicia).

2.2. Tarificación del seguro de automóvil

Una vez tratadas las coberturas y las garantías que proporcionan los seguros de automóvil, ahondaremos en el cálculo del precio de los seguros, el cual es denominado prima. Con respecto a esta prima, es importante tener en cuenta que se rige por el régimen de libre competencia, pero siempre respetando los principios de equidad y de suficiencia de la prima. Por otra parte, no se requiere la aprobación de la Dirección General de Seguros y Fondos de Pensiones para establecer las tarifas, las notas técnicas que las avalan ni los modelos de póliza, pero siempre deberán estar a disposición de dicha administración para su consulta.

En el precio de un seguro se debe incluir al menos la prima de riesgo, un recargo destinado a hacer frente a los gastos de gestión, administración y comercialización y una serie de comisiones externas establecidas por la normativa vigente, como son la contribución con el Consorcio de Compensación de Seguros y con el Fondo Nacional de Garantía. De esta manera, la ecuación a través de la cual se calcula una prima es la siguiente:

$$P'' = \frac{P}{1 - (g_d + g_i + g_b)}$$

Ecuación 1: prima de un seguro de automóvil

En la anterior ecuación, se pueden encontrar las siguientes variables:

- P'': prima de tarifa.
- P: prima pura.
- g_d: tanto por uno de recargo para gastos de consumo diferido.
- g_i: tanto por uno de recargo para gastos de consumo inmediato.
- g_b: tanto por uno de recargo para beneficio.

La prima pura la esperanza matemática de la siniestralidad y, además, la parte de la prima de tarifa (la prima que se cobra a los asegurados) que se utilizará para asumir los riesgos de los siniestros que puedan suceder. La diferencia entre la prima de tarifa y la prima pura serán aquellos recargos destinados a los gastos de la entidad y a los beneficios de la misma, los cuales son aplicados sobre la prima pura a través de las variables g_x.

Como es de imaginar, la verdadera dificultad del cálculo de esta cuota radica en sólo una de estas partes: en la prima pura. Es imprescindible realizar una estimación que sea suficiente para poder afrontar el pago de los siniestros de la compañía. Las aseguradoras actualmente se basan en estadísticas, con el objetivo de valorar el riesgo individual que cada póliza aporta a la cartera de seguros de la compañía.

A pesar de ello, la labor del cálculo de la prima de riesgo es especialmente difícil en el ramo de los seguros de automóvil debido a que la cantidad de factores que influyen en la ocurrencia de siniestros es muy elevada y no siempre se dispone de información real para analizarlos. Además, el enorme tamaño de las carteras de seguros hará que, sin duda, estos sean muy heterogéneos. Los expertos en seguros (actuarios) se enfrentan a un problema de muy difícil solución: distribuir el coste de los siniestros de una forma justa entre los distintos asegurados de la compañía, garantizando que cada uno de ellos pague en función del riesgo que aporte a la cartera.

Existen dos maneras de tarificar las pólizas, es decir, de determinar la prima, dentro de una cartera heterogénea: a priori y a posteriori.

Segmentación de riesgos (tarificación a priori)

Este método consiste en segmentar la cartera de riesgos en clases con mayores niveles de homogeneidad, basándose en ciertas variables de clasificación, de forma que todas las pólizas dentro de una misma clase paguen la misma prima. Los factores más comúnmente utilizados para realizar esta segmentación son relativos al vehículo (marca, modelo, potencia, color, peso…), al conductor (edad, sexo, antigüedad del permiso de conducir, estado civil…) y a la conducción (uso del vehículo –particular, profesional, de alquiler, etcétera– y características de la zona de circulación).

Sistemas bonus-malus (tarificación a posteriori)

Este método es necesario porque existen factores que afectan al riesgo de un seguro de automóviles y que son complicados de medir y, por tanto, de cuantificar. Estos factores son, por ejemplo, los reflejos del conductor, sus hábitos de conducción, conocimiento de las normas de circulación, etc. Debido a esto, aunque se haya separado en clases la cartera de seguros, estas siguen siendo heterogéneas. Por todo esto, a mediados de los años cincuenta surgió el sistema bonus-malus, también llamado tarificación basada en la experiencia, la cual ajusta la prima individual en función del historial de siniestralidad de cada asegurado.

Sin embargo, este sistema sólo puede considerarse una vez que se disponga de información sobre la experiencia real del asegurado. La primera vez que una persona contrata un seguro, la entidad aseguradora le asigna una prima a priori sin tener en cuenta la información real sobre siniestralidad, ya que ésta no existe todavía.

Este sistema consiste en una serie de bonificaciones o reducciones que se aplican sobre la prima de los asegurados cuando muestran una falta de siniestralidad de manera reiterada en su póliza. Por el contrario, se aplican penalizaciones a los asegurados que añadan siniestros a la cartera.

El sistema bonus-malus es ventajoso tanto para el asegurador, que es capaz de garantizar que las primas sean suficientes para asumir los riesgos derivados de la cartera, así como para el asegurado, el cual recibirá un precio justo por su seguro.

Para que este sistema funcione correctamente es necesario que exista una base de datos común en el que todas las aseguradoras pongan en común los historiales de siniestralidad de cada uno de los asegurados. Si esto no fuera así, los asegurados podrían cambiarse de una compañía a otra para evitar una subida de prima ocasionada por un siniestro reciente. Por ello, se ha creado el fichero histórico de siniestralidad de conductores (SINCO), al cual actualmente el 85% del mercado de aseguradoras está adherido.

Por lo general, en los seguros de automóvil se utiliza un método de cálculo de prima que combina ambos métodos: en primer lugar se realiza una tarificación a priori, y el valor obtenido se corrige con bonificaciones y penalizaciones generadas a través del sistema bonus-malus.

2.3. Gestión de siniestros

La gestión de siniestros es un proceso complicado que pretende esclarecer los hechos, causas y circunstancias del siniestro, además de evaluar el coste económico del mismo. La gestión de los siniestros tiene cuatro fases, las cuales se pueden observar en la siguiente página.

1. Declaración del siniestro: consiste en notificar a la aseguradora de la ocurrencia del siniestro de forma verbal o escrita. La aseguradora puede solicitar al asegurado toda la información y documentación que estime oportuna para gestionar el siniestro.
2. Tramitación del siniestro: la entidad aseguradora debe abrir un expediente del siniestro incluyendo al menos la siguiente información: el número del siniestro, la persona que ha recibido la declaración, la persona que tramita el siniestro, el número de póliza afectada, los datos del asegurado, las circunstancias del siniestro y las consecuencias del mismo. La compañía debe hacer una primera valoración del coste económico del siniestro. El siguiente paso es comprobar que el asegurado cumple todas las condiciones para que se tramite la indemnización, como por ejemplo el pago de la prima y que las circunstancias del siniestro estén cubiertos por la póliza.
3. Peritación del siniestro: la labor de valorar el coste del siniestro y la investigación del mismo puede ser encomendada a un profesional (perito) si la compañía lo cree conveniente. El perito, una vez estudie el siniestro, elabora un informe pericial en el que pronuncia sus conclusiones, del cual depende el coste del siniestro y el deber o no de la aseguradora de hacerse cargo del mismo. Si el asegurado no está de acuerdo con el informe, podrá reclamar por vía judicial.
4. Liquidación del siniestro: una vez valorado e investigado el siniestro, la entidad aseguradora decide si pagará o no la indemnización y en qué cuantía, pudiendo el asegurado reclamar por vía judicial si no queda conforme.

Dado el elevado número de seguros de automóviles existente, las compañías se ven obligadas a optimizar los procesos de gestión de los siniestros. Es esto lo que motiva la existencia del Convenio entre entidades aseguradoras de automóviles para la indemnización directa de daños materiales a vehículos (CIDE). Este convenio se aplicará sólo cuando haya dos vehículos implicados en un siniestro, y se firme una declaración amistosa de accidente por ambas partes.

En caso de siniestro, y cumpliendo los requisitos anteriores, la responsabilidad será aplicada según las Tablas de culpabilidad que contiene el CIDE o en función del Código de circulación. La entidad aseguradora del perjudicado reclamará a la entidad aseguradora del responsable por la vía más rápida posible, y esta última debe aceptar o rechazar el caso en un plazo máximo de 72 horas.

Para una mayor velocidad del pago de las indemnizaciones, será la entidad del perjudicado la que abonará las mismas. Posteriormente este coste será abonado por la aseguradora del responsable, pero el CIDE establece que se abonará el coste medio, haciendo una regulación para saldar las diferencias más adelante. De esta forma, se consigue ahorrar en gastos de gestión de todas las entidades aseguradoras.

Por último, es importante hablar del resarcimiento de daños corporales, los cuales son de muy difícil valoración. Para solucionar este problema, se han creado una serie de tablas que ayudan en el cálculo de dichas indemnizaciones, el *Sistema para la valoración del daño corporal en accidentes de tráfico* (comúnmente conocido como Baremo), que fue incorporado como Anexo a la *Ley sobre responsabilidad civil y seguro en la circulación de Vehículos a motor*, mediante el *Real Decreto Legislativo 8/2004, de 29 de octubre*. Este Baremo fomenta que en los siniestros con daños corporales similares las indemnizaciones sean casi idénticas.

Este Baremo ayuda a generar una indemnización dependiendo de las condiciones del siniestro:

- En caso de muerte, se tiene en cuenta la edad de la víctima, el número de beneficiarios así como su relación con la misma. De aquí surge una indemnización básica, que es modificada a la alza o a la baja por determinados factores. Algunos de estos factores son los ingresos de la víctima, la forma de participación de la víctima en el accidente, si falleciesen los padres de la víctima en el accidente, etcétera.
- En caso de incapacidad permanente, la indemnización se establece con un sistema de puntuación dependiendo de la gravedad de las secuelas y de la edad de la víctima. Partiendo de esta indemnización básica, se aplican una serie de correcciones similares a los del caso de muerte.
- En caso de incapacidad temporal, la indemnización básica se fija en función del tiempo que ha estado de baja la víctima, su edad y el salario mínimo interprofesional. Al igual que en los dos casos anteriores, se aplican factores de corrección, pero esta vez por los perjuicios económicos que haya podido sufrir la víctima.

2.4. El fraude en los seguros de automóvil

El fraude es definido como ≪el intento de ocultar circunstancias o de distorsionar la realidad para obtener un beneficio más allá de la justa compensación≫ en el libro de Guillén Estany et al. [5].

El fraude a las entidades aseguradoras es considerado un delito por la Ley ya que no sólo se perjudica a la compañía, sino también al resto de asegurados, provocando una subida de primas global para compensar las pérdidas ocasionadas por los defraudadores. Hay dos momentos clave en los que se puede producir el fraude: en el momento de la contratación y en la declaración de los siniestros.

En el primer caso, no existe un coste patrimonial directo para la empresa aseguradora, ya que éste no se verá reflejado hasta que ocurra un siniestro. Generalmente este tipo de fraude consiste en ocultación o falsedad de información a la hora de contratar una póliza con la finalidad de obtener una prima inferior. En ocasiones, se puede llegar a producir la contratación de una póliza

cuando el siniestro ya ha ocurrido sin notificar de ello a la compañía, para que sea ésta quien asuma dicho siniestro.

Con respecto al segundo caso, se suele dar cuando el asegurado solicita indemnizaciones que no le corresponden o declara unos daños superiores a los que realmente han ocurrido. Habitualmente este tipo de fraude se basa en simular la ocurrencia de un siniestro, falsear las circunstancias del mismo o abultar las consecuencias del mismo.

A lo largo de los años, se ha demostrado que la inversión de dinero en investigación de fraude ha demostrado reportar grandes ahorros de dinero a las entidades aseguradoras. La investigación de fraude se da principalmente de las siguientes maneras:

- Formación de peritos: este tipo de inversión tiene como su principal fin detectar el fraude de taller, el cual es uno de los fraudes más importantes en nuestro país.
- Formación de los tramitadores: invirtiendo en tramitadores se consigue que éstos detecten de manera más fácil y eficiente las incidencias o los casos anómalos excluidos de la cobertura de la póliza.
- Uso de inteligencia artificial: la informática está siendo cada vez más importante para agilizar y automatizar la investigación de siniestros. Estas herramientas informáticas, ayudadas por la estadística y la econometría permiten ayudar a las compañías a detectar fraude de manera precisa.

Sin embargo, a pesar de todas las medidas que se toman en España sigue existiendo fraude en los seguros de automóvil, y es necesario seguir creando y mejorando métodos de detección de este fraude.

2.5. Minería de datos

A día de hoy estamos abrumados con datos. La cantidad de datos que existen en este mundo y en nuestras vidas crece día a día, segundo a segundo, y no parece que vaya a parar de crecer nunca. Los computadores hacen muy fácil almacenar datos que en otros momentos de la historia simplemente se habrían desechado o ignorado. En este momento es muy barato comprar discos duros de enorme capacidad para almacenar datos, y dejar para más adelante la tarea de organizarlos y desechar los que no son de interés. Por otra parte, mientras navegamos por Internet en nuestro día a día, cada visita, cada operación, cada clic que hacemos queda almacenado, nada se desecha porque todo puede ser de interés para las empresas si es analizado adecuadamente.

Precisamente debido a este problema surge la minería de datos, la cual se define como el campo de las ciencias de la computación que busca encontrar patrones entre volúmenes enormes de datos. Con la minería de datos surge un nuevo potencial de los datos, al cual no se le había dado

demasiada importancia hasta el momento: el valor. Hernández Orallo, Ramírez Quintana y Ferri Ramírez [6] afirman que ≪los datos pasan de ser un "producto" […] a ser una "materia prima" que hay que explotar para obtener el verdadero "producto elaborado", el conocimiento […] ≫.

Se puede pensar que en este apartado se está realizando un cambio radical de tema, pero esto no es así puesto que la minería de datos está tremendamente relacionada con casi todos los aspectos de la vida cotidiana y, por tanto, con los seguros de automóvil. Es fácil imaginarse la cantidad de información que tendrán almacenada las empresas aseguradoras, talleres, empresas de peritaje, etcétera. Es entonces cuando entra en juego la minería de datos.

En muchas ocasiones el proceso de analizar los datos y extraer conclusiones es realizado de forma manual por seres humanos. Esto sucede, por ejemplo, en la investigación médica, donde grupos de médicos analizan en conjunto la evolución de enfermedades para obtener conclusiones como qué rangos de edad son más vulnerables, por qué medios se transmiten, etcétera. Sin embargo, esta forma de actuar es tremendamente lenta, cara y subjetiva. En algunos casos este método no es factible dada la cantidad de datos a analizar o las características de los mismos. La minería de datos busca resolver este tipo de problemas mediante el uso de computadores.

Por tanto, se puede afirmar que la minería de datos consiste en analizar grandes volúmenes de datos con el fin de descubrir información que está contenida en ellos, que es desconocida, pero que puede ser de utilidad, mediante el uso de procesos informáticas que consisten en la búsqueda de patrones comunes, relaciones o reglas que existan en los mismos, los cuales serán útiles para ayudar a explicar estos datos y poder realizar predicciones a partir de ellos.

La dificultad de este proceso no sólo reside en la complejidad de los algoritmos de minería de datos, sino que también en la eficiencia de estos procesos, ya que muchos análisis de minería de datos requieren una respuesta muy rápida, y no se puede esperar una gran cantidad de tiempo para obtener los resultados, ya que entonces estarán obsoletos. Un claro ejemplo de este problema es la publicidad en Internet. Empresas como Google, Amazon, eBay, entre otras muchas, muestran en su publicidad productos o servicios relacionados con la actividad que está realizando el usuario en las páginas que éste visita y las búsquedas que ha realizado con anterioridad. Recomendar al usuario el producto que busca en el momento adecuado puede resultar en una venta, mientras que si este proceso se demora mucho, cuando se presente al usuario la recomendación puede que éste ya haya adquirido otro producto.

La minería de datos se enmarca en un proceso que consta de cinco fases, como se puede observar en la siguiente ilustración.

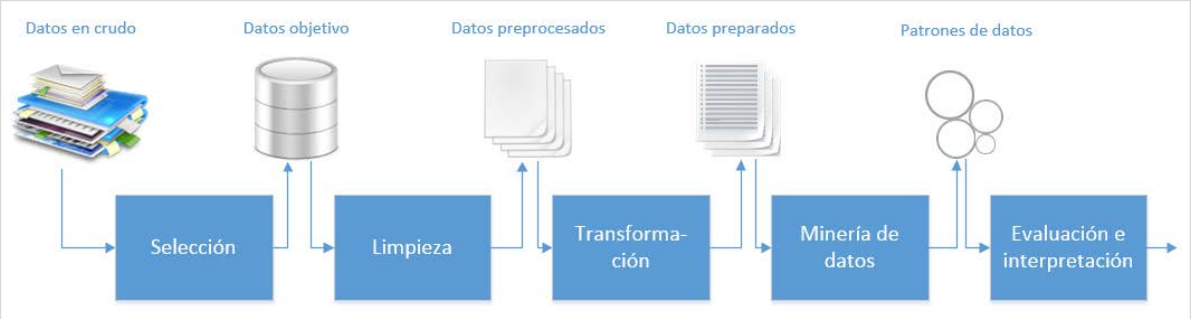


Ilustración 1: proceso de minería de datos

- Selección de los datos: se evaluará, a partir de los datos disponibles, las variables de inferencia y las variables de cálculo.
- Limpieza de los datos: se debe analizar de qué datos se puede extraer información útil y, para ello, preprocesar los datos eliminando datos irrelevantes, erróneos y atípicos.
- Transformación de los datos: se deben preparar los datos para el proceso de minería de datos, mediante técnicas de normalización, aleatorización, reducción de dimensiones y separación de conjuntos.
- Extracción del conocimiento (minería de datos): se debe aplicar uno de los procesos de minería de datos de los que se explicarán a continuación para obtener los patrones, reglas o relaciones deseadas.
- Evaluación e interpretación de los datos: esta fase consiste en analizar si los resultados obtenidos son coherentes y si se puede generar a partir de ellos un modelo final adecuado.

Es recomendable que este proceso se repita en varias ocasiones para obtener unos mejores resultados o perfeccionar el modelo resultante.

Existen dos tipos de objetivos existentes en la resolución de problemas de minería de datos. Estos objetivos pueden ser predictivos o descriptivos:

- Objetivos predictivos: los problemas se basan en la existencia de un conjunto de variables conocidas, a través de las cuales se pretende predecir el valor de una variable desconocida. Un ejemplo de este tipo de objetivos es la predicción de los intereses del usuario en plataformas publicitarias, para ofrecer productos relacionados con sus intereses. En este tipo de objetivos se enmarcan los problemas de clasificación y regresión.
- Objetivos descriptivos: en estos problemas lo que se pretende es la búsqueda de patrones comunes entre los datos a analizar, de forma que sea posible definir cada uno de estos patrones y comprender su naturaleza. Este tipo de objetivos es característico de los procesos de agrupamiento, las correlaciones y las reglas de asociación (incluidas las secuenciales).

Dentro de la minería de datos existen diversos tipos de tareas, cada una de las cuales se corresponde con un proceso que es aplicable a un determinado tipo de datos, o a obtener un determinado tipo de conclusiones.

Clasificación

Se trata de la tarea más común en la minería de datos. En esta tarea, se parte de una base de datos donde cada observación de la base de datos está asignada a una clase diferente. Esta clase es especificada mediante un atributo, que toma diferentes valores discretos, correspondiendo cada uno de ellos a una clase diferente, de manera similar a un número de identificación. El resto de atributos son utilizados para predecir la clase a la que pertenecen las nuevas observaciones que vayan apareciendo a lo largo del tiempo.

De una forma más técnica, el objetivo de este proceso es maximizar la precisión de la clasificación de nuevas observaciones. La razón de esta precisión es calculada como el cociente entre el número de predicciones correctas y el número de predicciones totales.

Además de la tarea de clasificación clásica, existen ciertas variantes de la misma, como el aprendizaje de rankings, el aprendizaje de preferencias, el aprendizaje de estimadores de probabilidad, etc.

Regresión

La regresión es una tarea que consiste en obtener una función que asigna a cada instancia un valor real. A diferencia de la clasificación, en este caso se trata de predecir un valor numérico. En este caso, el objetivo es minimizar el error entre el valor predicho y el valor real.

Agrupamiento (clústering)

El proceso de agrupamiento se basa en la obtención de grupos a partir de las observaciones de una base de datos. En este caso, a diferencia de la tarea de clasificación, no se trata de asignar las observaciones a clases sino de generar las clases en sí. Se busca maximizar la similitud de los datos contenidos dentro de un mismo grupo, mientras que la similitud entre los distintos grupos formados sea mínima.

Los grupos formados pueden ser o no disjuntos, es decir, puede o no haber observaciones que pertenezcan a distintos clústeres a la vez, dependiendo del estudio a realizar.

Correlaciones

Se trata de una tarea con la que se pretende comprender el grado de similitud entre los valores de dos variables numéricas. El método de medida de correlación más común es el coeficiente de correlación r, el cual es un valor comprendido en el intervalo [-1, 1], donde

cuanto mayor sea la cercanía al valor 1 o -1, las variables serán más correlacionadas positiva o negativamente. El valor cero indicará que no existe ningún tipo de correlación entre ellas. Este coeficiente se calcula a través de la siguiente fórmula:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

Ecuación 2: cálculo del coeficiente de correlación de Pearson

Reglas de asociación

Las reglas de asociación son muy similares a las correlaciones, ya que tiene como objetivo identificar relaciones entre atributos categóricos de la base de datos. La forma más común de este tipo de reglas consiste en buscar la relación de un atributo con otro, es decir, comprender qué valor toma en un atributo *X* dado el valor de otro atributo *Y*. Estas reglas no implican que exista una relación causa-efecto.

Esto sería utilizado, por ejemplo, por una tienda para identificar que dos productos se suelen comprar juntos, aunque estos no tengan ninguna relación entre sí, y podría ser utilizado no sólo para labores de marketing (sugerir al cliente adquirir un producto), sino también para labores logísticas (ordenar de determinada forma un almacén).

Existen también las reglas de asociación secuenciales, que son un caso particular de este tipo de reglas, que se utiliza para determinar patrones secuenciales en los datos. La diferencia con respecto a las reglas secuenciales son que en este caso se introduce una nueva variable: el tiempo.

2.6. Aprendizaje automático

El aprendizaje automático, también conocido por su nombre en inglés *Machine Learning*, es un campo de la informática que trata de estudiar y construir algoritmos que puedan aprender y realizar predicciones sobre datos. Estos algoritmos se basan en la construcción de modelos a partir de entradas de ejemplo, para poder realizar predicciones o decisiones guiadas por dichos datos, en lugar de seguir unas instrucciones estáticas plasmadas a priori en un programa informático.

El aprendizaje automático no sólo pretende construir estos algoritmos para los datos que han sido dados, sino que también pretende mejorar el rendimiento de la tarea de aprendizaje así como poder dar respuesta a situaciones desconocidas a partir de las observaciones previas.

El principal problema de esto, desde el punto de vista de la informática, es cómo saber si una máquina ha aprendido o no. Este problema no es tan fácil de resolver como en un centro

educativo, donde se hacen preguntas a los alumnos para medir los conocimientos que han adquirido.

La forma más común de comprobar el aprendizaje es considerar el rendimiento de un algoritmo a la hora de resolver un problema, es decir, observar el comportamiento y compararlo con el comportamiento pasado. Aunque esta definición del conocimiento es mucho más objetiva y parece ser más satisfactoria, el comportamiento, en muchos campos, cambia de manera que se obtiene mejor rendimiento en el futuro. Esto ocurre, por ejemplo, en el día a día de todos nosotros, al enseñar a un niño atarse los cordones de los zapatos, pues cada día tardará menos en hacerlo, pero no se producirá nuevo aprendizaje, ya que no han aumentado sus conocimientos sino su habilidad al atárselos o, lo que es lo mismo, su rendimiento. A este fenómeno se le llama entrenamiento.

Por tanto, el aprendizaje automático se basa en la obtención de habilidad para realizar una tarea de manera más eficiente gracias al aprendizaje.

Los dos tipos de aprendizaje automático más importante son el aprendizaje supervisado y el aprendizaje no supervisado.

2.6.1. Aprendizaje supervisado (clasificación)

Se trata de un método de aprendizaje automático que trata de inferir una función a través de unos datos, los cuales serán denominados datos de entrenamiento, de los que se conoce el resultado. Un escenario óptimo permitiría al algoritmo determinar correctamente el valor de nuevas observaciones no conocidas previamente.

El proceso del aprendizaje supervisado generalmente comienza haciendo un primer proceso proporcionando como entrada una serie de vectores que contiene los datos, así como el valor de salida deseado.

Para continuar, se proporciona un segundo grupo de vectores, de los que también se conoce el resultado. En este caso no se indica el resultado deseado al algoritmo sino que se utiliza para comprobar si el resultado obtenido es correcto. Este paso es muy importante para asegurarse de que no se realiza un sobreajuste de los datos, lo que generalmente está ocasionado por entrenar los datos de manera excesiva.

En la siguiente ilustración se puede observar una clasificación realizada en dos grupos (azul y rojo), la línea de división generada por el algoritmo de clasificación (verde) y la línea que sería óptima para realizar la clasificación (negro).

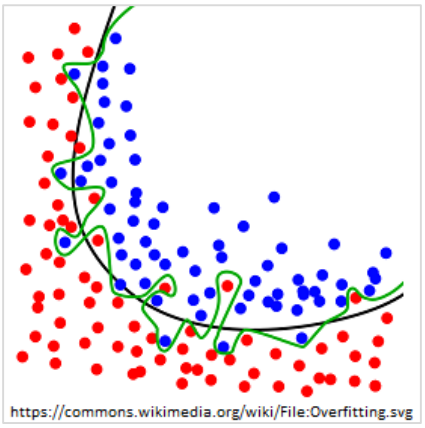


Ilustración 2: ejemplo de sobreajuste de los datos

Los algoritmos de clasificación generan un modelo a partir de los datos recibidos, pudiendo formar dos tipos distintos de modelos, conocidos como modelos de caja blanca y modelos de caja negra. Los modelos de caja blanca dan como resultado un algoritmo que el propio usuario puede interpretar sin problema, como por ejemplo un conjunto de reglas. En el caso de los modelos de caja negra, el algoritmo de predicción de la categoría es opaco al usuario, generalmente debido a su complejidad (por ejemplo, si para clasificar el elemento hay que dibujar un hiperplano).

Dentro de este tipo de aprendizaje se enmarcan los algoritmos de clasificación, siendo el más popular el árbol de decisión.

2.6.1.1. Árboles de decisión - Random Forest

Un árbol de decisión es una herramienta de ayuda a la toma de decisiones que utiliza un modelo en forma de árbol para tomar una decisión final. En el campo del aprendizaje automático, la decisión final consiste en la elección de un grupo al que pertenece la observación analizada. A continuación se puede observar un ejemplo de un árbol de decisión.

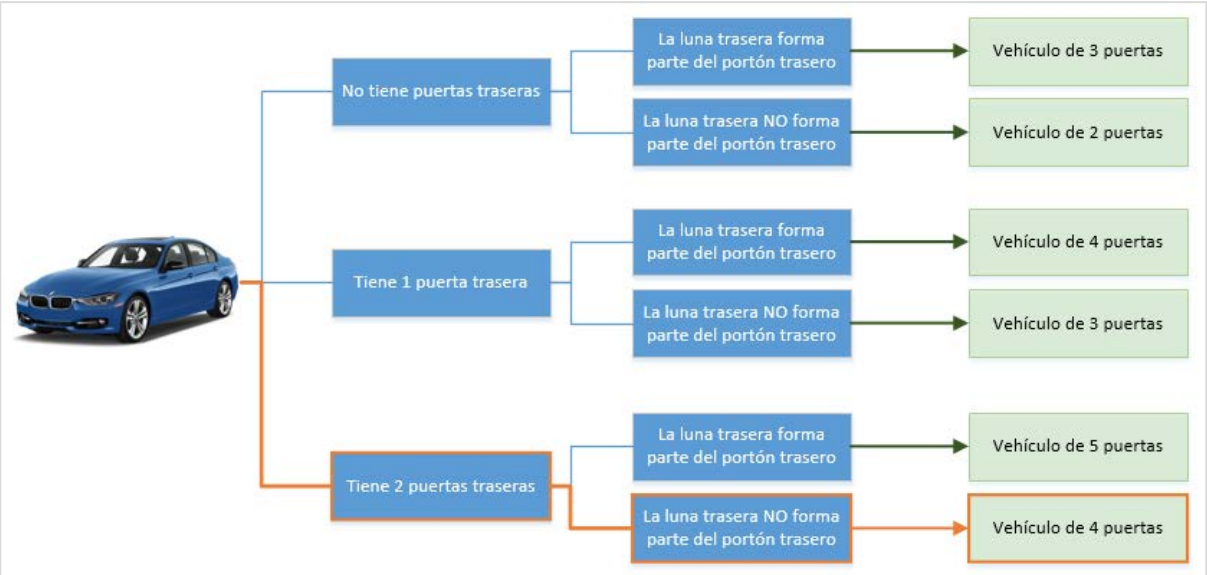


Ilustración 3: ejemplo de árbol de decisión

En este árbol de decisión podemos observar cómo se categoriza un vehículo según su número de puertas. En este caso, se transita entre las hojas del árbol de acuerdo a las características del vehículo (tiene dos puertas traseras y la luna trasera no forma parte del portón trasero), y se obtiene la categoría del vehículo (*Vehículo de 4 puertas*).

Los algoritmos de clasificación basados en árboles buscan generar este tipo de árboles a partir de una serie de datos previamente categorizados. Como se pudo observar, se trata de un algoritmo de caja blanca, ya que el algoritmo de predicción es conocido por el usuario.

Los algoritmos informáticos para crear árboles de decisión se basan en dos fases: entrenamiento y pruebas. Para realizar el proceso los datos de entrada son divididos en dos bases de datos, en una proporción definida por el usuario, de manera que se disponga de datos diferentes para cada una de las dos fases. La fase de entrenamiento consiste en la generación del árbol de decisión a través de casos conocidos, y la fase de pruebas consiste en comprobar cómo clasifica el árbol generado una serie de casos conocidos, comprobando si la categoría en la que la observación es clasificada coincide con la categoría real del caso de prueba.

El algoritmo *Random Forest* es uno de los más populares para generación de árboles de decisión con variables discretas en dos o más categorías. Este algoritmo consiste en crear una combinación de árboles predictores, de manera que cada uno de ellos depende de un vector aleatorio, basándose en una técnica llamada *Bootstrap Aggregating*. Esta técnica consiste *embolsar* las observaciones de la base de datos, es decir, empaquetar las observaciones de la base de datos en paquetes pequeños, todos ellos del mismo tamaño. De esta forma, los árboles es entrenado con estos pequeños conjuntos de entrenamiento, y el resultado es mejorado en gran medida.

2.6.2. Aprendizaje no supervisado (agrupamiento)

El aprendizaje no supervisado consiste en encontrar estructuras ocultas en datos sin etiquetar. Como no se conoce la estructura real existente entre estos datos, no hay ninguna señal que indique el error o acierto al evaluar una nueva situación.

El método más popular de aprendizaje no supervisado es el agrupamiento o clustering. Se trata de una técnica de minería de datos cuyo fin es formar de manera automática grupos de elementos, denominados clústeres, basado en una medida de similitud o distancia. El objetivo de los algoritmos de clústering consiste en que los grupos formados tengan una similitud entre sí baja (*similitud inter-clúster*), pero que los elementos de cada uno de ellos tengan una similitud alta (*similitud intra-clúster*).

A continuación se analizarán en detalle los dos algoritmos de clústering que resultan más relevantes para este estudio, así como los algoritmos de medida de distancia más populares.

2.6.2.1. Canopy

Este proceso utiliza una medida de distancia rápida con ayuda de dos umbrales definidos por la persona que ejecuta el algoritmo, *U1* y *U2*, para realizar una separación de los datos en clústeres en una única pasada sobre los datos. El proceso empieza con una base de datos y una lista vacía de grupos. En cada iteración, toma un vector de la base de datos y añade un nuevo grupo con el vector extraído como centro. Para continuar, tomará uno a uno el resto de valores de la base de datos, calculando la distancia que tiene cada uno de ellos a los centros de los grupos formados hasta el momento. Si la distancia es inferior o igual a *U1*, lo añadirá al grupo. Si la distancia es inferior o igual a *U2*, es eliminado de la lista y se previene que sea elegido como centro en el resto de iteraciones.

Para que la comprensión de este algoritmo sea más sencilla, se ha realizado un diagrama de flujo que explica al detalle su funcionamiento.

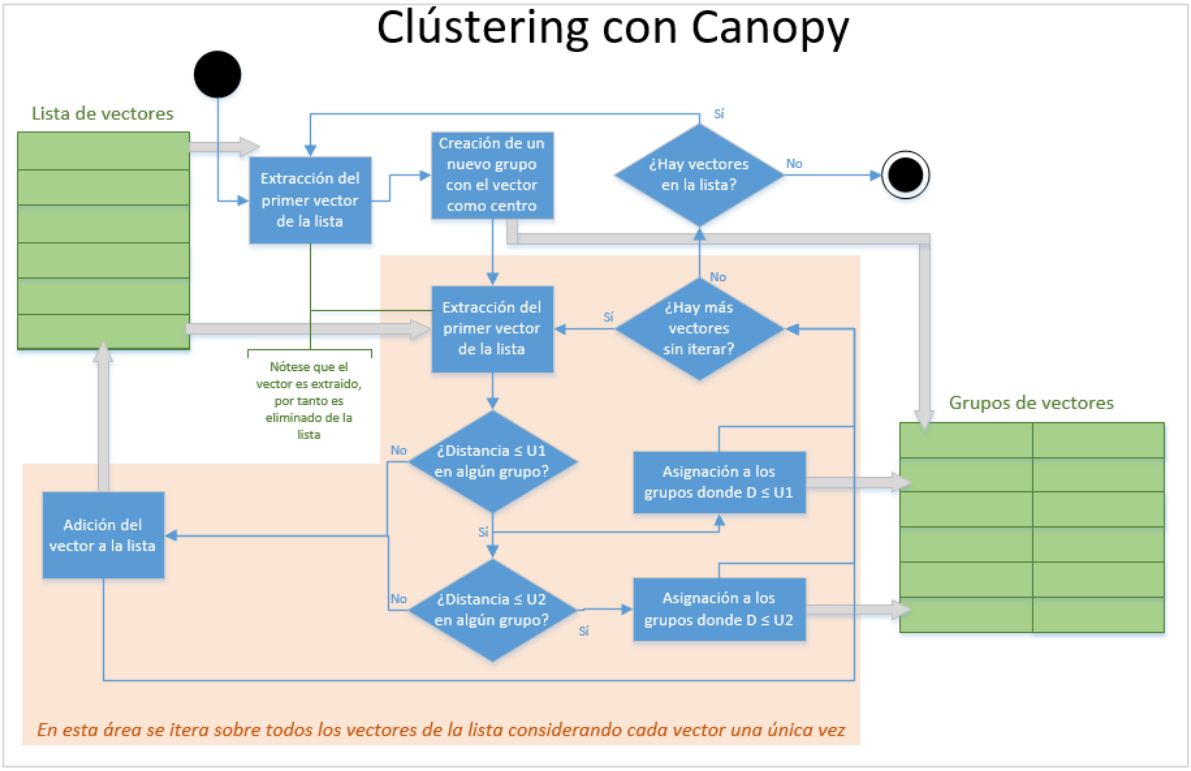


Ilustración 4: algoritmo de clústering Canopy

Se trata de un algoritmo que es la evolución del algoritmo *Expectation Maximization* (*EM*), cuya única diferencia con Canopy es que con *EM* los puntos que no están dentro de un clúster afectan de manera muy elevada a los elementos de los clúster que están en formación durante el proceso. El proceso Canopy simplifica en gran medida el algoritmo *EM* porque en cada iteración, en lugar de calcular la distancia de un vector a todos los puntos de los grupos, sólo la calcula con respecto al centroide de los mismos.

2.6.2.2. K-Means

El método k-Means, también conocido por su menos utilizada traducción al español *k-Medias*, es una técnica de cuantificación vectorial que es comúnmente utilizado para procesos de agrupamiento o clústering en el ámbito de la minería de datos.

La aplicación más habitual de este algoritmo utiliza una técnica de refinamiento iterativo, que consiste en que en cada iteración sobre el conjunto de datos se trata de mejorar la aproximación a la solución. Este algoritmo, dado un vector inicial de k centroides $m_1^{(1)}, m_2^{(1)}, \dots, m_k^{(1)}$, procesa los datos alternando entre dos etapas:

- **Etapa de asignación**: se asigna cada observación al grupo cuya media sea más cercana (de acuerdo al algoritmo de medida de distancia que se esté empleando).

$$S_i^{(t)} = \left\{ x_p : (x_p - m_i^{(t)})^2 \leq (x_p - m_j^{(t)})^2 \ \ \forall j, 1 \leq j \leq k \right\}$$

Ecuación 3: etapa de asignación en el algoritmo k-Means

  En esta ecuación, x_p representa cada uno de los vectores de la base de datos a analizar, k representa el número de clústeres que se desean obtener y $m^{(t)}$ se corresponde a la media del centroide en la iteración t. Cada uno de los vectores x_p puede ser asignado a un único clúster ($S_i^{(t)}$) aunque pudiera ser asignado en dos o más de ellos.

- **Etapa de actualización**: se calculan las nuevas medias, que pasarán a ser los centroides de las observaciones en los nuevos clústeres, de acuerdo a la siguiente fórmula:

$$m_i^{(t+1)} => \frac{1}{\left|S_i^{(t)}\right|} \sum_{x_j \in S_i^{(t)}} x_j$$

Ecuación 4: etapa de actualización en el algoritmo k-Means

De esta forma, se continúa alternando entre estas dos etapas, hasta que una iteración de este proceso no se mueva ningún vector entre los grupos, es decir, hasta que una iteración no ocasione ningún cambio en los mismos.

El vector inicial al que se ha hecho mención, el cual es necesario para llevar a cabo el proceso, se suele elegir de manera aleatoria entre todos los vectores de la base de datos a analizar. Una buena elección de estos vectores puede hacer el proceso más corto, pero no debería afectar al resultado final del mismo, ya que la repetición de las iteraciones acaba distribuyendo los clústeres siempre de una manera muy similar.

Para facilitar la comprensión de este algoritmo de clústering, se ha diseñado un diagrama de flujo que explica paso a paso cómo funciona este algoritmo, el cual se puede observar en la siguiente página.

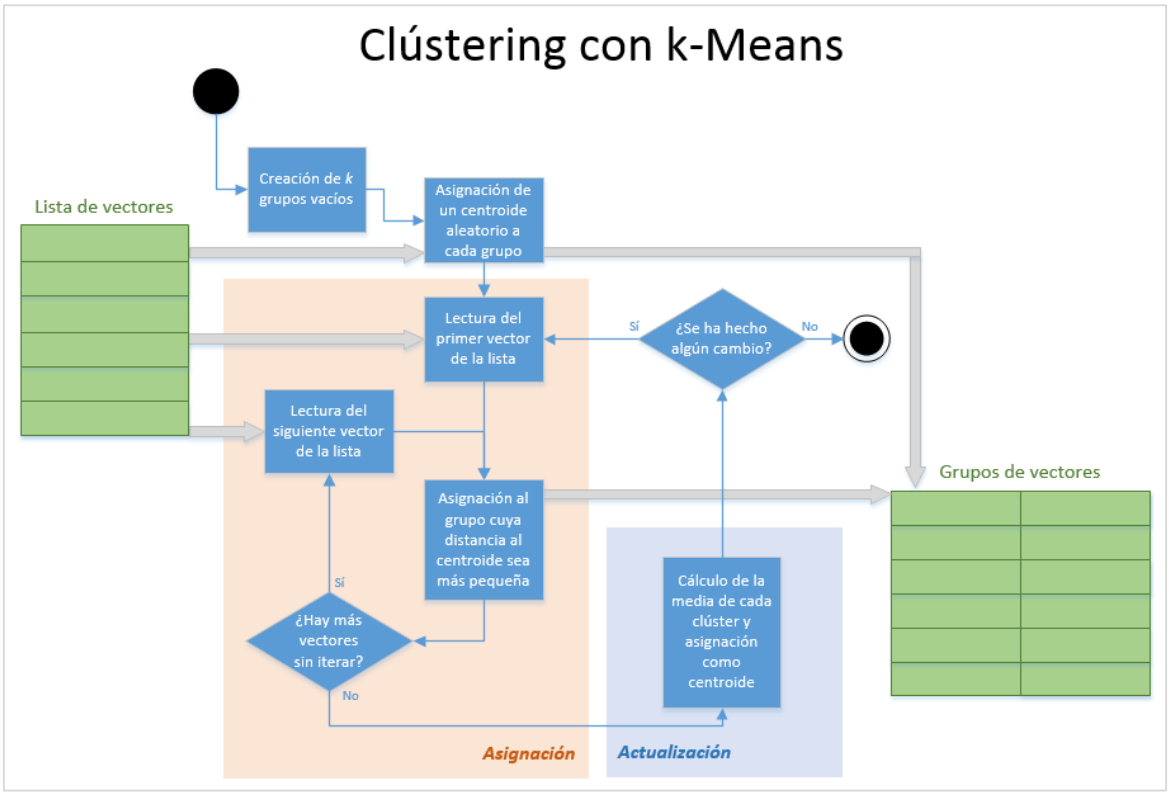


Ilustración 5: algoritmo de clústering k-Means

2.6.2.3. Algoritmos de medida de distancia

Una vez se ha definido el objetivo del análisis y cómo llevarlo a cabo, es necesario pensar en cómo medir la similitud de las observaciones contenidas en la base de datos. Esto es un factor decisivo a la hora de realizar el análisis, ya que determina la forma en la que se compararán dos elementos de una base de datos, y por tanto influye directamente en el resultado final del clústering. De esta forma, un algoritmo de medida de distancia nos indicará el grado de disimilitud, es decir, cómo de diferentes son dos observaciones a comparar.

La necesidad de obtener esta información da lugar a numerosos algoritmos de medida de distancia, donde cada uno de ellos será útil para un tipo determinado de datos. A continuación se explicarán aquellos algoritmos que se consideran más relevantes para este estudio.

Distancia Euclídea

La distancia euclídea es la medida de distancia más frecuentemente utilizada, dado que es muy intuitiva y fácil de comprender. Este algoritmo se basa en el Teorema de Pitágoras para calcular la distancia entre dos vectores $\vec{x}$ e $\vec{y}$, como se puede observar en la siguiente ecuación:

$$d(\vec{x},\vec{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

Ecuación 5: cálculo de Distancia Euclídea

Esta fórmula no es más que el cálculo de la longitud de la línea recta que conecta a los dos vectores que se están estudiando en un espacio n-dimensional.

Distancia Euclídea al Cuadrado

Se trata de una variante muy utilizada de la distancia euclídea. Este algoritmo se corresponde exactamente con el cuadrado del resultado de la distancia euclídea. A continuación se puede observar la fórmula utilizada por este algoritmo para comparar dos vectores $\vec{x}$ e $\vec{y}$:

$$d(\vec{x},\vec{y}) = (x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2$$

Ecuación 6: cálculo de Distancia Euclídea al Cuadrado

De esta forma, se consigue aplicar progresivamente más peso a la distancia cuanto más elevada es ésta, penalizando a los vectores cuanto más lejos están entre sí.

Distancia Manhattan

La distancia Manhattan, también conocida como Geometría del taxista, se basa en sumar las diferencias de cada una de las coordenadas cartesianas de los vectores a comparar. A través de la siguiente ecuación se calcula la distancia entre dos vectores $\vec{x}$ e $\vec{y}$:

$$d(\vec{x},\vec{y}) = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_n - y_n|$$

Ecuación 7: cálculo de Distancia Manhattan

Este algoritmo de medida de distancia es más fácil de entender si se dibuja sobre una cuadrícula y se compara con la distancia euclídea:

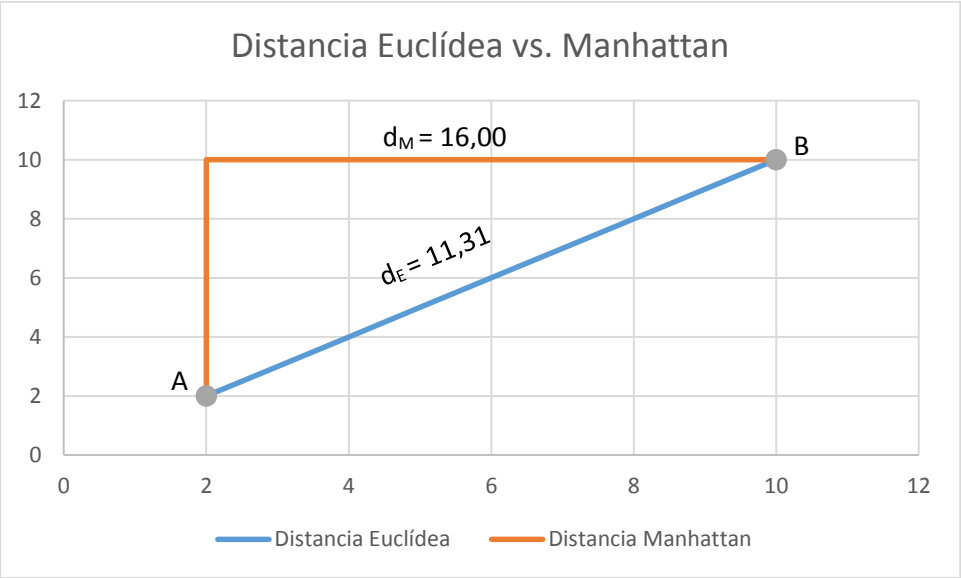


Ilustración 6: Distancia Euclídea vs. Distancia Manhattan

Distancia Mahalanobis

Este algoritmo de medida de distancia compara la distancia entre un par de puntos y una distribución, en lugar de comparar únicamente los dos puntos entre sí, como se hace con la distancia euclídea. Se trata de una generalización que mide la cantidad de desviaciones típicas que separan el punto de la media de la muestra. Para calcular la distancia Mahalanobis, en primer lugar se calcula la matriz de coeficientes S y se utiliza la siguiente fórmula:

$$d(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T * S^{-1} * (\vec{x}-\vec{y})}$$

Ecuación 8: cálculo de Distancia Mahalanobis

En esta ecuación $\vec{x}$ e $\vec{y}$ representan cada una de las dos observaciones a comparar de forma vectorial. La variable S^{-1} representa la matriz de covarianza invertida. Aunque el resultado de esta operación sea un vector, se puede tomar la distancia como la suma de cada uno de los valores del mismo.

La contrapartida de este algoritmo de medida de distancia es que requiere una cantidad de memoria muy elevada para almacenar la matriz de coeficientes, o si es almacenada en disco producirá mucho tráfico y ralentizará el proceso.

Distancia del Coseno

Este algoritmo de medida de distancia mide el ángulo existente entre dos puntos. Cuando este ángulo es pequeño, significa que los vectores apuntan hacia el mismo lugar, y en algunas bases de datos esto significa que los puntos son cercanos. Esto es llevado a cabo a través de la siguiente ecuación, donde p y q son los dos vectores a comparar:

$$d(\vec{x},\vec{y}) = 1 - \frac{(x_1y_1 + x_2y_2 + \cdots + x_ny_n)}{\left(\sqrt{(x_1^2 + x_2^2 + \cdots + x_n^2)}\sqrt{(y_1^2 + y_2^2 + \cdots + y_n^2)}\right)}$$

Ecuación 9: cálculo de Distancia del Coseno

Distancia Chebyshev

Este algoritmo de medida de distancia, también conocida como métrica máxima o métrica L_∞, define la distancia entre dos vectores como la más grande de las diferencias existentes entre cada una de las coordenadas.

$$d(\vec{x},\vec{y}) = \lim_{k\to\infty}\left(\sum_{i=1}^{n}|x_i - y_i|^k\right)^{\frac{1}{k}}$$

Ecuación 10: cálculo de Distancia Chebyshev

Una forma sencilla de comprender este algoritmo es interpretar esta distancia como la cantidad de movimientos que tendría que hacer un rey, para moverse desde uno de los vectores hasta el otro, en un tablero de ajedrez. En la siguiente ilustración se puede observar un rey de ajedrez y la distancia Chebyshev que tiene cada una de las casillas hasta el mismo.

Ilustración 7: distancia Chebyshev representada en un tablero de ajedrez

2.7. Estudios y proyectos previos

Se ha realizado una búsqueda muy extensa para tratar de encontrar otros estudios relacionados con lo que se va a tratar en este proyecto, y únicamente se ha encontrado un estudio muy similar al que aquí se va a llevar a cabo.

El trabajo de fin de grado llamado *Análisis de datos aplicado a siniestros de automóviles* [7], fue publicado en Junio de 2015 por Eduardo González González en el repositorio de documentos de la Universidad Carlos III de Madrid. Este estudio pretende ≪realizar un modelado de las características esenciales que muestra un siniestro de automóvil […] con el fin de automatizar el proceso de tasación y facilitar la detección de valoraciones materiales fraudulentas o erróneas≫. Como se puede observar, la finalidad de este proyecto es bastante similar a lo que se va a abordar en este estudio.

E. González realiza una separación en clústeres por severidad y zonas de impacto de la misma manera que se va a realizar en este estudio, de una forma menos profunda que en este estudio, ya que sólo realiza una clasificación de primer nivel. Adicionalmente, explica el proceso de creación de árboles de decisión utilizando como partida los grupos que ha generado. Para todo esto utiliza el software WEKA, el cual, como se describirá más adelante, tiene diversas carencias dado que sólo se puede ejecutar en una máquina y es muy lento.

En este proyecto se ampliar todo lo que E. González realiza en su trabajo, partiendo de la misma base de datos, pero empezando de cero utilizando con un software que permita realizar el proceso de manera distribuida. De esta forma, se tratará de realizar un análisis de rendimiento de estas tecnologías y además mejorar el rendimiento y la precisión de los resultados obtenidos. Adicionalmente, se realizará una agrupación de segundo nivel (se subdividirán los grupos en grupos más pequeños) y se utilizarán algoritmos de clasificación más complejos.

Por otra parte, la directiva Solvencia II, a la que se hizo referencia en la introducción, define unos requisitos de capital en función del riesgo asumido por las entidades aseguradoras, de manera que se reduzca el riesgo de insolvencia. Por ello, esta normativa explica la necesidad imperiosa de mejorar los métodos de cálculo de las primas a cobrar a los asegurados, para poder cumplir estos requisitos de capital repartiendo las primas de una forma adecuada y justa entre cada uno de los asegurados de las compañías.

3. Diseño de la solución

En este apartado se detallarán todas aquellas decisiones que han sido tomadas para llevar a cabo este estudio, de manera debidamente justificada. Estas decisiones engloban el software, el hardware y los algoritmos que se utilizarán y la forma en la que éstos interactuarán entre sí. Adicionalmente, se detallan las alternativas de que se dispone y se exponen las razones por las cuales se ha decidido no utilizarlas. Por último, también se describirá, paso por paso, el proceso que se va a llevar a cabo para llevar a cabo el estudio.

3.1. Herramientas utilizadas

Para el desarrollo de este estudio, se han utilizado diversas herramientas de software. Se ha intentado en todo momento que el software utilizado sea libre y código abierto, especialmente aquel que es más importante para el estudio.

3.1.1. Apache Mahout 0.9

Ilustración 8: logotipo de Apache Mahout

En primer lugar, el pilar básico de esta investigación será el entorno de desarrollo de algoritmos de aprendizaje automático Mahout. Esta herramienta ha sido creada por la fundación Apache para la ejecución de este tipo de algoritmos de manera que sean escalables.

Esta herramienta ha sido seleccionada por diversos motivos, entre los cuales el más importante es la facilidad a la hora de escalar problemas con él, gracias a la integración nativa del mismo con herramientas de cómputo distribuido como Hadoop y Spark. Otra de las grandes ventajas del mismo es que es de código abierto y está en continuo desarrollo, por lo que de esta forma estaremos seguros de que el software está probado suficientemente y es estable.

La versión de Mahout elegida para la resolución de este problema es la 0.9, liberada en enero de 2014. A pesar de que existen versiones más recientes, no es recomendable utilizarlas dado que no están adecuadamente documentadas ni probadas. De hecho, los manuales online y en papel más recientes sólo documentan hasta esta versión, y no existe ninguno que trate sobre las versiones 0.10 ni 0.11.

Esta librería no sólo nos permitirá interactuar con ella a través del terminal de Linux, sino que también dispone de una API en Java que nos permite desarrollar aplicaciones completas que lo integren. Para utilizar las API de Mahout será necesario utilizar un IDE (generalmente Eclipse o Netbeans) con el *plugin* de Maven, el cual es un potentísimo gestor de dependencias.

La alternativa más importante a Mahout es WEKA (*Wakiato Environment for Knowledge Analysis*), la cual es una herramienta desarrollada por la Universidad de Waikato (Nueva Zelanda). Se trata una colección de algoritmos de aprendizaje automático que, igualmente que Mahout, pueden ser invocados directamente o a través de una API de Java.

A pesar de que este software es muy potente, no ha sido elegido porque sus algoritmos son menos potentes que los de Mahout y requieren más tiempo de procesamiento. En el trabajo de E. González [7] se explica que dado el elevado tiempo de procesamiento de este software, tuvo que utilizar únicamente una parte de los datos, ignorando gran parte de los datos existentes en la base de datos. A diferencia de Mahout, WEKA no es compatible con software de procesamiento distribuido, por lo que se ha descartado la opción de utilizarlo. Sin embargo, en este proyecto se pretende hacer posible analizar bases de datos independientemente de su tamaño, considerando para los análisis todos los datos contenidos en la misma.

Otras herramientas como CLUTO (University of Minnesota), CLUSTER 3.0 (Stanford University) o PYCLUSTER (University of Tokyo) muestran las mismas carencias que WEKA, por lo que tampoco han sido consideradas para este estudio.

3.1.2. Apache Hadoop 1.2.1

Ilustración 9: logotipo de Apache Hadoop

Para realizar la distribución del proceso entre varios computadores, dada la restricción que establece Mahout, sólo podíamos escoger entre Hadoop, Spark y Hive. Se ha determinado que la mejor de las tres alternativas es Hadoop porque es la herramienta que mejor integrada está con Mahout, y lleva formando parte del mismo desde las primeras versiones. La versión elegida es la 1.2.1 porque es la versión más alta recomendada por Apache para la versión de Mahout elegida.

Adicionalmente, resulta de especial interés el entorno MapReduce, que forma parte del núcleo de Hadoop y está pensado para desarrollar aplicaciones que tratan cantidades muy grandes de información en un número elevado de nodos en un entorno fiable y tolerante a fallos. Los trabajos MapReduce se basan en separar una base de datos en segmentos independientes que son procesados por las tareas *map* de una forma completamente paralela. El entorno ordena la salida de los mapas y son utilizadas como entrada para las tareas *reduce*, las cuales se encargan de combinar los resultados obtenidos.

Los procesos MapReduce con Hadoop están formados por un único `JobTracker` situado en el computador maestro, el cual se encarga de monitorizar el proceso global, y un `TaskTracker` en cada uno de los nodos esclavos, el cual monitoriza cada tarea independientemente. Si hay algún fallo en el desarrollo de las tareas, el maestro asignará dicha tarea de nuevo a alguno de los otros esclavos. Habitualmente, el computador maestro funciona como esclavo de sí mismo, asegurándose de que se aprovecha al máximo el rendimiento del clúster.

Hadoop, además, requiere una infraestructura muy simple para funcionar correctamente, pues las máquinas únicamente deben estar conectadas por red, independientemente de que sea cableada o inalámbrica. Tanto el sistema de ficheros distribuido como el procesamiento de manera distribuida se realizan utilizando esta tecnología. Es por ello que para incrementar el rendimiento de Hadoop conviene que todos los equipos que se utilicen estén conectados en la misma conexión de área local, para reducir la latencia de las conexiones y aumentar la velocidad de las mismas.

La única alternativa a Hadoop soportada por Mahout es Apache Spark. Se trata de un motor de procesado de datos a larga escala. Spark es un proyecto muy joven, cuya primera versión estable fue lanzada en Mayo de 2014. La característica más importante de Spark es el soporte a conjuntos de datos residentes distribuidos (*Resilient Distributed Dataset* – RDD). Los RDD son colecciones de elementos particionados a lo largo de los nodos de un clúster que puede operar de forma paralela. Otra característica muy importante es la posibilidad de uso de variables compartidas entre todos los nodos que ejecutan operaciones en paralelo.

Esta alternativa, a pesar de mostrar un mejor rendimiento en sistemas paralelizados que Hadoop, todavía no está muy probada ni muy documentada, además de que la integración con Mahout está todavía en fase experimental, por lo que podría llevar a errores en el estudio a realizar. De cualquier manera, puede ser muy interesante considerar su uso cuando éstas sean más estables, ya que se espera que den un rendimiento muy superior.

3.1.3. Eclipse Mars

Ilustración 10: logotipo de Eclipse

Se ha elegido este entorno de desarrollo Java para desarrollar la solución de este problema, ya que es un entorno fácilmente integrable con Maven a través del plugin *m2e*. Este entorno, además, puede integrarse de una manera sencilla con herramientas de gestión de versiones como git, de

forma que se lleve un seguimiento completo de los cambios que se realizan en el código del programa. Esto además nos permitirá detectar cualquier fallo rápidamente dado que este seguimiento de los cambios nos permite observar qué ha cambiado con cada nueva funcionalidad que se ha desarrollado, ahorrando tiempo de rastreo de errores.

Eclipse, además, permite generar ficheros JAR ejecutables de una forma muy sencilla, incluyendo en su interior todas las dependencias necesarias para su funcionamiento, tanto las generadas por Maven como las introducidas manualmente.

3.1.4. Matlab R2015a

Ilustración 11: logotipo de Matlab

Matlab es un potentísimo entorno que, entre otras muchísimas funciones, incluye herramientas de cálculo avanzado y representación de gráficas en 2 y 3 dimensiones sin largas esperas. De esta forma, será muy fácil conseguir analizar los resultados obtenidos por Mahout, creando pequeños ficheros de Matlab que realicen representaciones leyendo los ficheros en CSV.

Este programa, además, tiene una documentación extensísima, que en conjunto con la cantidad de información al respecto en Internet, facilita mucho su uso y hace posible que se avance a mayor velocidad para hacer gráficas y tratar archivos.

Adicionalmente, un factor decisivo en la elección de utilizar Matlab ha sido que está disponible para ser utilizado en las aulas de la Universidad, sin necesidad de realizar el gran desembolso que implica adquirir una licencia.

Otras alternativas gratuitas como Octave, además de estar menos documentadas, no daban un rendimiento tan elevado como Matlab. Sin embargo, son muy recomendadas en caso de que no se disponga de una licencia de Matlab, ya que son capaces de llevar a cabo las mismas tareas a pesar de su inferior rendimiento.

Se ha descartado la utilización de Microsoft Excel porque, tras varias pruebas de rendimiento, con bases de datos de gran número de entradas el consumo de memoria RAM y procesamiento de Excel es muy elevado, y produce numerosos cuelgues en los computadores. Sin embargo, Matlab no da ninguno de estos problemas.

3.1.5. Otro software

Ilustración 12: logotipos de Notepad++, Excel, PowerPoint y Visio

Por último, se han utilizado otras herramientas de software para tareas de menor relevancia en el estudio, pero resultan de especial ayuda para el desarrollo ágil del mismo.

El potente programa de hojas de cálculo Microsoft Excel 2013 permite realizar estadísticas sobre los diferentes ficheros CSV de gran tamaño que se tratarán a lo largo de los diferentes análisis de este estudio. A diferencia de las alternativas de otras organizaciones, como LibreOffice Calc, Excel maneja las hojas de cálculo de gran tamaño con mucha más velocidad y sin grandes tiempos de espera.

Para el tratamiento rápido de ficheros de texto se ha utilizado Notepad++, el cual dispone de herramientas de búsqueda y reemplazo con expresiones regulares muy potente, además de interpretar correctamente los saltos de línea de sistemas Linux a diferencia del Bloc de notas tradicional de Microsoft.

La redacción del presente documento se ha realizado con el programa Microsoft Word 2013 y las diferentes gráficas han sido trazadas con ayuda de Microsoft PowerPoint 2013 y Microsoft Visio 2013.

3.1.6. Sistemas Operativos

Ilustración 13: logotipos de Ubuntu, Xubuntu y Windows

Para llevar a cabo estas tareas, en primer lugar es necesario analizar qué sistemas operativos se van a utilizar. Dadas las distintas elecciones de software que se han realizado, es necesario seleccionar un sistema operativo compatible con todo este software, que permita desarrollar el estudio de la manera más eficiente posible.

Para realizar el proceso con Mahout, Hadoop y Eclipse, se utilizará el sistema operativo Ubuntu, que dada su gran popularidad es muy estable y tiene un consumo reducido de recursos. Además,

todo el software que se va a utilizar o bien está previamente compilado o bien está probado extensivamente en este sistema operativo.

Por otra parte, los análisis con Matlab y el software de Microsoft se deben realizar en equipos con sistema operativo Windows, así que se utilizarán equipos con Windows en sus versiones 8.1 y 10 para tal efecto.

3.2. Descripción de los datos

Se dispone de una base de datos de peritación de siniestros correspondiente a un vehículo de una marca y modelo concretos. Estos datos han sido tomados durante un periodo de 3 años, y están relacionados con siniestros que han ocurrido en territorio español. El vehículo de estudio es un turismo de categoría B, de tamaño reducido que es comúnmente conocido como utilitario o compacto. Este vehículo dispone de dos versiones, una con 3 y otra con 5 puertas (incluyendo el portón trasero entre ellas en ambos casos). Sin embargo, en la base de datos no existe ninguna variable que permita distinguir entre ambos modelos. En total se dispone de 329.013 partes de accidente correspondientes a este automóvil.

Los datos de que disponemos se encuentran en formato CSV, es decir, se encuentran en un fichero de texto en el que los datos son representados en forma de tabla. En este fichero, cada línea representa un siniestro, y cada columna (separadas entre sí por comas) representa cada uno de los atributos del siniestro.

Cada uno de los siniestros representados en este fichero contiene 719 atributos que lo definen detalladamente. Estos atributos contienen los siguientes datos:

- Número de secuencia (único por cada parte de siniestro).
- Número de historia (número de veces que se ha modificado el parte).
- Mapa de bits de órdenes de pintura sobre las piezas (de entre un total de 237 piezas), donde 0 significa que no ha sido pintada y 1 que sí lo ha sido.
- Mapa de bits de órdenes de reparación sobre las piezas (de entre un total de 237 piezas), donde 0 significa que no ha sido reparada y 1 que sí lo ha sido.
- Mapa de bits de órdenes de sustitución sobre las piezas (de entre un total de 237 piezas), donde 0 significa que no ha sido sustituida y 1 que sí lo ha sido.
- Número de piezas que han sido sustituidas.
- Número de piezas que han sido reparadas.
- Coste total de la mano de obra.
- Coste total de la pintura.
- Coste total de las piezas sustituidas.

- Coste total del siniestro (suma de los tres costes anteriores). Este coste supone el precio que la aseguradora ha tenido que desembolsar por el siniestro.

Antes de continuar, será necesario aclarar lo que significan varios conceptos, para comprender perfectamente qué es lo que nos ofrece esta base de datos.

El primer lugar, se considera como orden de pintura aquella reparación que requiera únicamente trabajos de pintura o pulido de una pieza. Este tipo de órdenes originan en su mayor parte los costes de pintura.

Con respecto a las órdenes de reparación, son aquellas que son causadas por una deformación o rotura de la pieza dañada, y requieren mano de obra para reparar el desperfecto. En este caso las órdenes de reparación originan, principalmente costes de mano de obra.

Por último, las órdenes de sustitución son aquellas que surgen cuando una pieza ha quedado tan deteriorada que es más barato reemplazarla que repararla, y conlleva el reemplazo de la pieza dañada por una completamente nueva. Este tipo de órdenes generan principalmente costes de sustitución de piezas, así como costes de mano de obra.

En el Anexo II se puede observar de forma detallada la descripción de cada uno de los atributos disponibles, así como de las piezas contempladas en esta extensa base de datos.

3.3. Elección de la arquitectura

Un paso crucial para llevar a cabo este estudio es elegir de forma adecuada una arquitectura que permita la ejecución de los algoritmos de manera que se hallen los resultados de forma eficiente. A continuación se describirán todos los sistemas informáticos de los que se dispone y se realizarán diversas pruebas de rendimiento que permitan elegir la mejor combinación de equipos y conexiones para poder desarrollar este estudio.

3.3.1. Hardware disponible

Para llevar a cabo esta investigación, se dispone de una potente arquitectura informática cedida por el Grupo de Inteligencia Artificial Aplicada (GIAA) de la Universidad Carlos III de Madrid. Esta arquitectura está formada por dos computadores domésticos y dos supercomputadores. Además de estos cuatro computadores se incluye un quinto computador privado para completar la arquitectura completamente. Las características de los mismos se pueden observar en la siguiente página.

	Computador 1	Computador 2	Computador 3
Nombre de la máquina	giaa-edge1	giaa-edge2	slave3
Modelo del procesador	Intel Xeon E5-2609	Intel Xeon E5-2609	Intel Core i5-2400
Número de procesadores	2	2	1
Número de núcleos por procesador	4	4	4
Velocidad de reloj	2,4GHz	2,4GHz	3,1GHz
Módulos de memoria	4	4	2
Cantidad de memoria total	64GB	64GB	8GB
Velocidad de memoria	1.600MHz	1.600MHz	1.067MHz
Sistema Operativo	Ubuntu Server 14.04 LTS	Ubuntu Server 14.04 LTS	Xubuntu 15.04
P.V.P. Aproximado	3.800€	3.800€	550€

Tabla 1: computadores utilizados para el proceso (I)

	Computador 4	Computador 5
Nombre de la máquina	slave4	slave5
Modelo del procesador	Intel Core 2 Quad Q9300	Intel Core i7-3632QM
Número de procesadores	1	1
Número de núcleos por procesador	4	4
Velocidad de reloj	2,5GHz	2,2GHz
Módulos de memoria	2	2
Cantidad de memoria total	6GB	8GB
Velocidad de memoria	667MHz	1.333MHz
Sistema Operativo	Xubuntu 15.04	Xubuntu 15.04
P.V.P. Aproximado	480€	650€

Tabla 2: computadores utilizados para el proceso (II)

Nótese que el precio de venta al público (P.V.P.) ha sido estimado utilizando los precios de venta al público, los cuales están disponibles en las páginas web www.pccomponentes.com y www.dell.com, y fueron consultados el día 20/06/2015. En caso de que éstos estuvieran obsoletos se ha estimado el precio utilizando los precios de componentes de características similares. Estos datos serán utilizados únicamente para verificar en los siguientes apartados la relación del precio con el rendimiento obtenido por cada uno de los computadores.

Existe, además, un sexto computador, el cual también es privado, pero sus características no son de importancia para el estudio dado que no realizará ningún análisis crucial, sino que se utilizará para representaciones gráficas, análisis estadísticos y redacción de documentos.

3.3.2. Pruebas de rendimiento

Una vez definido el software que se va a utilizar, es indispensable analizar el rendimiento del mismo en los distintos computadores que nos han sido cedidos para asegurarnos de un rendimiento óptimo en los distintos análisis que se van a realizar.

Se han conseguido diversas bases de datos de acceso libre en Internet para realizar este tipo de pruebas. En esta tabla se pueden observar todas las bases de datos utilizadas para hacer las pruebas de rendimiento de los computadores con el entorno MapReduce de Hadoop. La finalidad de estas pruebas es comprobar qué factores afectan en el rendimiento del mismo, por lo que se utilizarán distintas bases de datos, con distintas características de manera que, con cada una de ellas, se extraigan conclusiones muy diferentes. Las características de las bases de datos utilizadas se pueden utilizar a continuación:

	Base de datos 1	Base de datos 2	Base de datos 3	Base de datos 4
Nombre	Gutenberg Ebooks	Household Power Consumption	Partes de accidente	Bag of Words
Autor	Project Gutenberg	University of California, Irvine	*Confidencial*	University of California, Irvine
Contenido	Palabras	Mediciones de consumo energético de hogares	Partes de accidente de vehículos	Apariciones de palabras en diferentes textos
Observaciones	2.283.212	2.075.259	329.013	483.450.157
Variables por observación	1	7	719	3
Tipo de variables	Palabras	Numéricas	Numéricas	Numéricas
Tamaño	6,62MB	86,90MB	460,16MB	7,27GB
Proceso a realizar	WordCount	k-Means	k-Means	k-Means

Tabla 3: bases de datos utilizadas para realizar pruebas de rendimiento

Nótese que la base de datos 3 se corresponde con la base de datos que utilizaremos para este estudio y que ha sido descrita en el apartado 3.2.

Antes de continuar, es preciso detallar en qué consisten los procesos *WordCount* y *k-Means* que se han mencionado en la tabla 3:

- WordCount: este proceso consiste en contar el número de apariciones de cada palabra en un conjunto de textos. Este proceso representa la salida como un documento de texto mostrando en cada línea cada una de estas palabras junto al número de ocasiones en las que aparece. Este proceso es interesante porque a priori parece que nos permitirá medir el rendimiento de un clúster con Hadoop en un proceso sencillo.

- k-Means: se trata de un método de minería de datos que permite el agrupamiento de un conjunto de observaciones en distintos grupos o clúster de características similares. Este método ha sido explicado en detalle en el apartado 2.6.2.2, ya que es uno de los métodos más importantes de este estudio. Utilizar este método para probar el rendimiento nos ayudará a analizar en el caso concreto de nuestro problema cómo se comporta el clúster de computadores.

Previamente a la muestra de estos datos, es necesario conocer las distintas arquitecturas utilizadas para realizar estas pruebas. Estas arquitecturas han sido detalladas en la siguiente tabla.

Configuración	Maestro	Esclavos	Observaciones
1	Computador 3	Computador 3	*Arquitectura formada por un único computador corriente que realiza el proceso por sí mismo.*
2	Computador 3	Computador 3 Computador 4	*Arquitectura formada por dos computadores corrientes que realizan el trabajo en conjunto.*
3	Computador 3	Computador 3 Computador 4 Computador 5	*Arquitectura formada por un clúster de 3 computadores corrientes que realizan el proceso en conjunto.*
1s	Computador 1	Computador 1	*Arquitectura formada por 1 supercomputador que realiza todo el proceso por sí mismo.*
2s	Computador 1	Computador 1 Computador 2	*Arquitectura formada por 2 supercomputadores realizando el trabajo en paralelo.*
2s+2	Computador 1	Computador 1 Computador 2 Computador 3 Computador 4	*Arquitectura formada por 2 supercomputadores y 2 computadores corrientes realizando el trabajo en paralelo.*

Tabla 4: arquitecturas utilizadas para realizar pruebas de rendimiento

A continuación se detallarán los resultados obtenidos con cada una de estas bases de datos procesada por cada una de las arquitecturas descritas anteriormente, y se obtendrán conclusiones que servirán para decidir cuál de estas arquitecturas es más conveniente para realizar este estudio.

En las siguientes gráficas se puede observar el rendimiento obtenido al ejecutar las pruebas con todas las bases de datos especificadas previamente. En el eje vertical se mostrará el tiempo medio de ejecución, en minutos, tras la repetición de la misma prueba en 10 ocasiones. Por otra parte, en el eje horizontal se mostrarán los computadores utilizados para realizar la prueba. Dada la

imposibilidad de representar los resultados de todas las bases de datos en la misma escala, se han decidido separar los resultados en dos ilustraciones distintas.

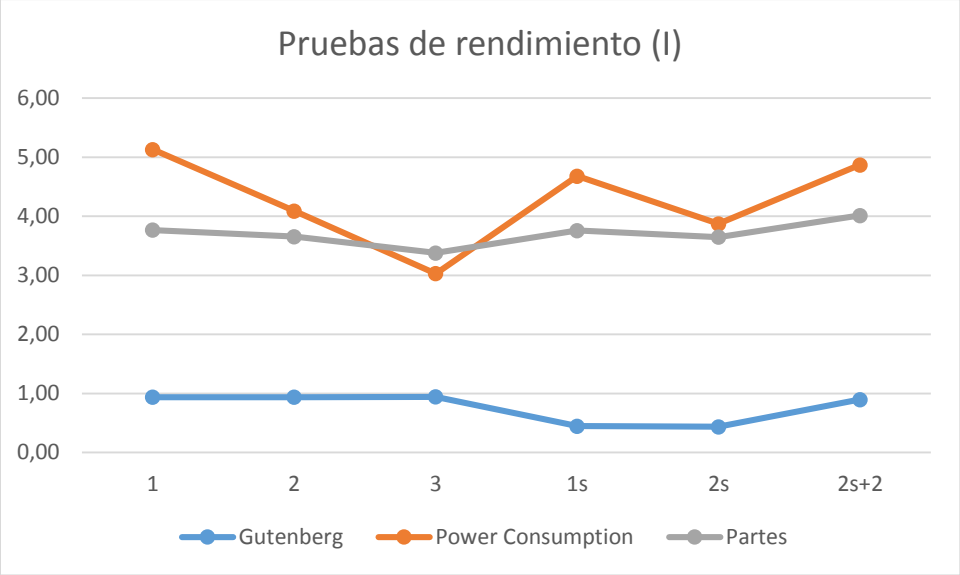


Ilustración 14: pruebas de rendimiento de la arquitectura (I)

Del gráfico anterior se pueden obtener numerosas conclusiones. En primer lugar, como era de esperar, la base de datos *Gutenberg* es demasiado pequeña como para plantearnos una arquitectura con Hadoop. De hecho, el hecho de paralelizar el proceso no reporta ninguna mejora en este caso, dando resultados en tiempos casi idénticos. Esto se puede observar tanto con computadores normales (arquitecturas 1, 2 y 3) como con supercomputadores (arquitecturas 1s y 2s). Como conclusión se puede extraer que, para este tipo de bases de datos, el hecho de disponer de un computador con mayor rendimiento reducirá el tiempo de cómputo notablemente; en este caso del 52.5%.

Con respecto a la base de datos *Power Consumption*, las conclusiones que se pueden extraer son ligeramente diferentes a las de la base de datos anterior. En este caso vemos como la adición de computadores similares al clúster reduce el tiempo de cómputo de manera lineal, ahorrando entre un 15 y un 25% de tiempo de cómputo con cada máquina similar añadida al clúster. En este caso, se comprende que esto sea así porque el procesamiento de los datos es complejo, al haber un gran número de observaciones, cada iteración del proceso será lenta al tener muchos datos que procesar, aunque no se realizarán muchas iteraciones dado que la cantidad de variables de cada observación no es muy elevada. Adicionalmente, se ha observado que la sustitución de computadores domésticos por supercomputadores no reporta una gran mejora en el tiempo de procesamiento de los datos, por lo que no parece razonable realizar una inversión en supercomputadores para analizar una base de datos similar a la que hemos utilizado en este caso.

Por otra parte, la base de datos de partes de accidente (*Partes*) muestra unas características diferenciadoras con respecto a la segunda base de datos. En este caso, la cantidad de observaciones es grande también, pero no tanto como en el primer caso, pero cada una de estas observaciones tiene cientos de variables. En este caso, se observa como la introducción de más máquinas o el uso de supercomputadores no aporta casi ahorro temporal al proceso. Esto se explica porque al tener tantas variables, pero una cantidad de observaciones más reducida, cada una de las iteraciones del proceso *k-Means* se ejecuta con gran velocidad, pero se realizan muchas iteraciones (entre 80 y 100 en cada ejecución). Como las iteraciones deben ser realizadas secuencialmente, es indiferente que haya más computadores en el clúster, porque estas iteraciones son de muy rápida ejecución.

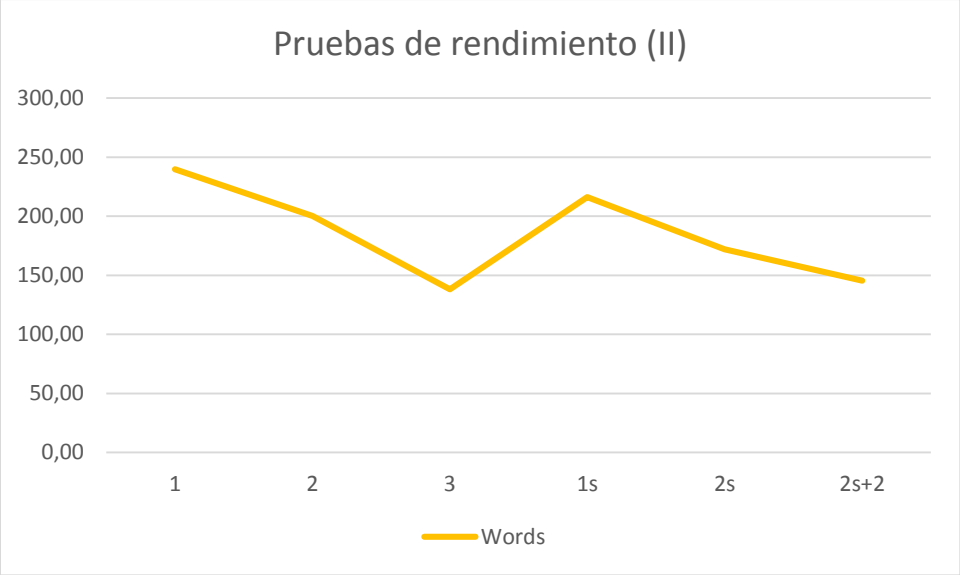


Ilustración 15: pruebas de rendimiento de la arquitectura (II)

Por último, la base de datos *Words* demuestra de nuevo que la adición de más computadores en un problema complejo mejora notablemente el tiempo de ejecución del algoritmo. En este caso, la mejora relativa es de entre el 15 y el 40%.

Por otra parte, en ambas ilustraciones hemos podido observar como la arquitectura *2s+2*, en todos los casos, es contraproducente. Parece que los computadores más pequeños actúan como cuello de botella en la arquitectura. Esto puede ser explicado fácilmente teniendo en cuenta que Hadoop separa la tarea en trozos del mismo tamaño, haciendo que todos los computadores tengan que procesar fragmentos del mismo tamaño sin tener en cuenta el tamaño total del fichero de entrada. Es por esto por lo que con ficheros de un tamaño no muy elevado, añadir más nodos a un clúster hará que el fichero se divida en más trozos y tarde más en ser procesado.

3.3.3. Conclusiones de las pruebas realizadas

Tras analizar todos los resultados obtenidos en el apartado anterior, podemos sacar varias conclusiones que nos ayudarán a tomar una decisión con respecto a la arquitectura a elegir para realizar el estudio.

La primera conclusión clara del estudio es que debe tenerse una base de datos muy grande y debe realizarse un proceso complejo para plantearse una arquitectura con Hadoop. El tamaño de la base de datos debería estar a partir de decenas de millones de datos, independientemente de la cantidad de columnas que exista en los mismos. Asimismo, es necesario también que se trate de un proceso complejo, como el algoritmo *k-Means* que hemos estudiado anteriormente. En procesos simples como *WordCount* será preciso disponer de bases de datos de un tamaño muy superior para que merezca la pena tener en cuenta Hadoop.

Por otra parte, se ha podido observar que realizar una inversión en grandes computadores no parece merecer la pena para llevar a cabo este proceso. En el caso de las pruebas realizadas con anterioridad, hemos podido ver como un clúster de tres computadores domésticos da resultados en menos tiempo que un clúster de dos supercomputadores. Con el coste de un único supercomputador podríamos adquirir aproximadamente siete computadores domésticos, los cuales darían un rendimiento mayor que los supercomputadores.

Para continuar, hemos visto que debemos elegir adecuadamente el número de nodos para realizar el estudio, pues elegir un número excesivo de nodos puede ser igual de negativo que elegir un número muy reducido de nodos. Es recomendable, para este tipo de prácticas, realizar una serie de pruebas previamente al estudio, tratando de verificar qué número de nodos es conveniente.

Por último, como se ha observado mezclar computadores de características muy diferentes puede generar cuellos de botella, por lo que es muy recomendable que todos los computadores del clúster sean de características similares para garantizar que el rendimiento sea óptimo.

3.3.4. Arquitectura elegida

Como vimos en el apartado anterior, la máquina más lenta de la arquitectura reduce la velocidad del proceso, ya que funciona como un cuello de botella, llegando incluso a hacer que el proceso sea igual o más lento que sin incluir dicha máquina.

Adicionalmente, como se ha visto al hacer las pruebas de rendimiento con la base de datos 3 (*Partes*), el tiempo de ejecución de este algoritmo es muy reducido y apenas varía al modificar el número de equipos o al cambiar entre computadores normales y supercomputadores.

Por esta razón, se ha decidido que, para el análisis de esta base de datos, se crearán tres clústeres independientes, de manera que permitan realizar varios análisis simultáneamente y obtener más

resultados en menos tiempo. Cada uno de los supercomputadores formará un clúster por sí solo, mientras que los computadores 3 y 4 formarán un clúster de dos nodos. De esta manera se obtendrá un rendimiento similar en los tres clústeres.

Para reducir el consumo de memoria y procesamiento, se ha determinado que en los computadores 1 y 2 se ha instalado el sistema operativo Ubuntu Server, el cual carece de interfaz gráfica, lo cual minimiza el consumo de memoria RAM y procesamiento de estos computadores, y así se podrá disponer de la máxima capacidad de procesamiento para esta investigación. La versión elegida es la 14.04, la cual es la última versión con soporte a largo plazo de Ubuntu Server. Esta versión tiene como principal ventaja la gran estabilidad del sistema así como el gran foco que se pone en la seguridad comparado con las versiones "estándar".

Sin embargo, los equipos 3 y 4 sólo se utilizarán además para monitorizar el proceso y consultar y visualizar los resultados ya que se encuentran en un entorno más favorable para su acceso y la conexión de pantallas, dado que los supercomputadores están instalados en un *rack* para servidores. Por tanto, se ha instalado en los computadores 3 y 4 el sistema operativo Xubuntu en su última versión (15.04), la cual dispone de una interfaz gráfica de reducido consumo de recursos (XFCE). De esta forma podremos disfrutar de las ventajas de disponer de interfaz gráfica especialmente para el análisis de los resultados, sin entorpecer la eficiencia del seguimiento del proceso de agrupamiento en tiempo real.

A pesar de que no se utilizarán todos los computadores en paralelo, se seguirá empleando una red privada para comunicar estos computadores entre sí. Esto es debido a que será habitual la transferencia de ficheros entre ellos para su análisis y almacenamiento. Esta es la razón por la que estos cuatro equipos se conectarán entre sí a través de una red de área local (LAN) creada con ayuda de un conmutador de red (conocido como switch), de forma que se eviten las demoras y retrasos propios del uso de conexiones a través de Internet para comunicarse entre sí.

Cuando se desea paralelizar utilizando Hadoop, también es muy recomendable utilizar una red de área local, dado que de esta forma se conseguirá que la comunicación entre ellos sea muy rápida y, lo que es más importante, que la latencia de las conexiones entre ambos será muy reducida. La latencia es un detalle muy importante dado que las comunicaciones que existirán entre ambos computadores serán muy frecuentes, porque el computador maestro necesita conocer el estado de su esclavo en cada momento.

3.4. Algoritmo distribuido

Para resolver este problema, se ha decidido implementar un sistema en dos fases que permita realizar el proceso de clústering de la manera más precisa posible en un tiempo razonable. Las dos fases a las que se ha hecho mención son las siguientes:

1. Estimación del número de grupos a través del algoritmo *Canopy*.
2. Separación de los siniestros en distintos grupos utilizando el algoritmo *k-Means*.

En los siguientes apartados se describirá detalladamente cómo se realizará este proceso y los factores que han sido tenidos en cuenta.

3.4.1. Estimación del número de grupos (Canopy)

Para la estimación del número de grupos se utilizará el algoritmo *Canopy* de Mahout. Este es el único algoritmo de agrupamiento de Mahout que no requiere conocer el número de grupos a priori, sino que el número es obtenido por él mismo. De esta forma será la opción perfecta para estimar el número de grupos en el que debemos separar el proceso.

Para ejecutar este algoritmo, Mahout requiere que se especifiquen los siguientes parámetros:

- *Fichero de entrada*: se trata de la base de datos en formato secuencial que se desea analizar. Debe contener como columnas las distintas variables de la base de datos y como filas las observaciones. Cualquier variable o dato que se no se desee tener en cuenta debe ser eliminado previamente. Este fichero debe encontrarse en el sistema de ficheros distribuido creado por Hadoop.
- *Umbral 1*: este umbral, que coincide con el umbral U1 definido anteriormente, establece cuál es la distancia máxima a la que puede estar un vector al centroide de un grupo para poder ser agregado al mismo.
- *Umbral 2*: este umbral, que coincide con el umbral U2 definido anteriormente, que debe ser siempre inferior al primero, establece la distancia máxima entre un vector y un centroide para que el primero sea considerado parte del clúster que se forma en torno al segundo y no se forme un nuevo clúster con él.
- *Algoritmo de medida de distancia*: Mahout incorpora diversos algoritmos de medida de distancia entre vectores a los que se puede acceder a través de su API Java. Estos algoritmos, algunos de los cuales fueron explicados en detalle con anterioridad, están incorporados en el paquete `org.apache.mahout.common.distance`, entre los cuales podemos encontrar *Chebyshev*, *Coseno*, *Euclídea*, *Mahalanobis*, *Manhattan*, *Minkowski*, *Euclídea al cuadrado*, *Tanimoto*, además de otras pequeñas variantes de estos. Sin embargo, se puede implementar cualquier medida de distancia en Java sin ningún problema y utilizarla con Mahout sin ningún problema.
- *Número máximo de iteraciones*: número máximo de veces que se puede iterar sobre la base de datos (de acuerdo a la definición del algoritmo *Canopy*).
- *Directorio de salida*: directorio en el que escribir los clústeres resultantes del proceso (en el sistema de ficheros distribuido.

La elección de estos tres argumentos será realizada en cada uno de los análisis de manera cuidadosa dependiendo de la naturaleza de los datos que se deseen analizar.

El algoritmo *Canopy* en sí es un algoritmo de agrupamiento que realiza una separación en clústeres, pero se trata de un algoritmo bastante anticuado que la fundación Apache no recomienda utilizar, dado que es muy sensible a los umbrales de los que se habló anteriormente. Sin embargo, lo mantienen en las nuevas versiones de Mahout porque es una opción perfecta para realizar una estimación del número de grupos que se desean crear, eligiendo cuidadosamente los valores para los umbrales y los algoritmos de medida.

3.4.2. Separación de los siniestros en grupos (k-Means)

Para llevar a cabo la separación en clústeres se empleará el algoritmo k-Means, que viene incluido en Mahout. Este algoritmo, como se trató con anterioridad en el apartado 2.6.2, realiza una separación en grupos de manera muy eficiente, ya que partiendo del número de grupos que se desean obtener, realiza la distribución de las observaciones en ellos de una forma muy precisa.

Este algoritmo requiere que se le proporcionen diversos argumentos:

- *Fichero de entrada*: al igual que en el caso anterior, se trata de la base de datos en formato secuencial que se desea analizar en la que sólo se incluyan los valores que quieran ser tenidos en cuenta. Este fichero debe encontrarse dentro del sistema de ficheros distribuido.
- *Directorio de clústeres iniciales*: k-Means requiere que existan unos grupos iniciales para realizar el proceso. En este caso se utilizarán los clústeres resultantes del proceso *Canopy* (deben encontrarse en el sistema de ficheros distribuido).
- *Algoritmo de medida de distancia*: se debe indicar un algoritmo de medida de distancia incluido en el paquete `org.apache.mahout.common.distance` o cualquier otro que cumpla las características requeridas por la API de Mahout como se indicó con anterioridad.
- *Número de grupos*: se debe indicar el número total de clústeres que se desea obtener con el proceso de agrupamiento.
- *Número máximo de iteraciones*: número máximo de veces que se puede iterar sobre la base de datos (de acuerdo a la definición del algoritmo *k-Means*).
- *Directorio de salida*: directorio en el que escribir los clústeres resultantes del proceso (dentro del sistema de ficheros distribuido).

Tras el proceso, Mahout creará un directorio llamado `clusteredPoints` dentro del directorio de salida, que contendrá un fichero con todas las observaciones de la base de datos así como la clave del clúster al que han sido asignados. Mahout escribirá este fichero exactamente en el mismo

orden que el fichero de entrada, por lo que será fácil relacionar ambos ficheros, lo cual es muy útil si se han eliminado variables que no querían ser tenidas en cuenta para el proceso.

3.4.3. Creación de un árbol de decisión (Random Forests)

Mahout contiene una implementación del algoritmo *Random Forests*, el cual fue descrito previamente, en la clase `org.apache.mahout.classifier.df.mapreduce.BuildForest`. El proceso de entrenamiento implementado por Mahout requiere los siguientes parámetros:

- *Fichero de entrada*: ruta a un fichero en formato CSV (debe estar en el sistema de ficheros distribuido creado por Hadoop), utilizando la coma como elemento para delimitar las diferentes columnas.
- *Número máximo de iteraciones*: número máximo de veces que se puede iterar sobre la base de datos de entrenamiento. Un valor elevado dará resultados más precisos pero reducirá el rendimiento del proceso.
- *Descripción de las columnas*: este parámetro permite especificar una ruta (relativa al sistema de ficheros distribuido) donde se encuentra la descripción delas columnas generada con la herramienta `org.apache.mahout.classifier.df.tools.Describe`.
- *Directorio de salida*: ruta relativa al sistema de ficheros distribuido creado por Hadoop donde se desea que se almacene el modelo creado por Mahout.

Una vez terminado el proceso, Mahout creará un fichero donde tendrá toda la información necesaria para poder realizar la clasificación de nuevas observaciones, codificado en un formato propio. Mahout proporcionará únicamente cierta información sobre los árboles generado, con información sobre el número de nodos, la media de profundidad de los árboles y la media de amplitud de los mismos.

3.5. Descripción del proceso de agrupamiento

Se ha creado un pequeño programa en línea de comandos en Java que puede ser invocado a través de la línea de comandos y que realiza el proceso de agrupamiento, también conocido por el término en inglés “clustering”, utilizando la API de Mahout. Este programa, además del proceso de agrupamiento, se encargará del preprocesado y postprocesado de los datos. Este programa realizará todo el proceso en función de ciertos parámetros que le sean proporcionados. En esta sección estudiaremos en detalle las tres fases del análisis, así como de la futura interpretación de los mismos con Matlab y Excel.

3.5.1. Preprocesado de datos

La fase de preprocesado pretende limpiar los datos de valores erróneos o extremos, así como adaptar los mismos a los requisitos de Mahout para poder procesarlos. El proceso de preprocesado consta de 6 etapas: limpieza de datos erróneos, eliminación de valores atípicos,

aislamiento de datos a analizar, normalización de los datos, copia del fichero al sistema de ficheros distribuido y conversión a fichero secuencial de acceso aleatorio. El funcionamiento de cada una de estas etapas se puede observar a continuación.

3.5.1.1. Limpieza de datos erróneos

En un fichero de datos tan largo y complejo como el que nos ocupa, es habitual que existan errores en alguna de las observaciones. Estos errores suelen consistir en errores tipográficos, o pequeñas incoherencias, y suelen estar originados por fallos humanos a la hora de introducir los datos en el sistema informático.

Para realizar cualquier tipo de análisis en estos datos el primer paso es eliminar del fichero de datos aquellas observaciones en las que se detecten estas incoherencias. Se han eliminado todas aquellas observaciones que cumplan al menos uno de los siguientes requisitos:

- Cuando exista algún valor de un mapa de bits de piezas de pintura, reparación o sustitución que no sea un 0 o un 1.
- Cuando exista algún coste de reparación, pintura o sustitución de piezas negativo.
- Cuando el coste total del siniestro no se corresponda con la suma del coste de reparación con el de pintura y sustitución de piezas.

3.5.1.2. Eliminación de valores atípicos

En cualquier análisis de este tipo la eliminación de datos atípicos es una parte esencial del preprocesado de una base de datos.

Un valor atípico es una observación que no sigue el patrón de distribución de un conjunto de datos, es decir, es una observación lejana al resto de los datos. En la ilustración 16 se puede observar un ejemplo de un valor atípico en un gráfico en dos dimensiones, donde el punto rojo se aleja claramente del patrón de los demás puntos de la muestra.

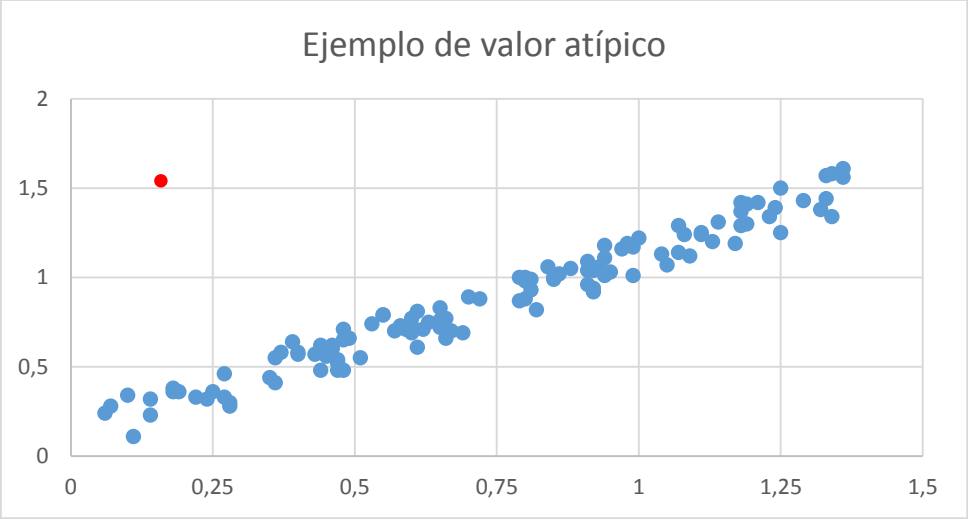


Ilustración 16: ejemplo de valor atípico

La existencia de valores atípicos puede darse como resultado de diversos factores, como por ejemplo errores humanos al transcribir los datos o acontecimientos fuera de la normalidad. Dado que estos valores atípicos no son valores habituales en la muestra, podrían alterar los resultados si se consideran como datos normales. Estos casos se pueden considerar casos extremos y excluirlos del análisis para evitar que un suceso que ocurre con muy poca frecuencia altere los resultados.

Mahout no incluye ningún proceso de eliminación de datos atípicos, así que se ha desarrollado un algoritmo propio que realiza un análisis sobre los datos y elimina aquellos que son atípicos. Este proceso consta de los siguientes pasos:

1. Se calcula la mediana de cada una de las variables de la base de datos. No se utiliza la media dado que ésta se ve afectada por los valores atípicos, pudiendo llegar a hacernos eliminar valores que realmente no son atípicos. En la siguiente ilustración se puede ver el efecto de un valor atípico en la media:

Ilustración 17: ejemplo de cómo afecta un valor atípico a la media de una muestra

   Se creará una nueva observación ficticia, estableciendo en cada variable el valor de las medianas recientemente calculadas.

2. Para cada valor de la base de datos, se almacenará el valor de la distancia de cada observación con respecto a la mediana almacenada en el paso anterior. Se utilizará para ello la distancia utilizando el método Mahalanobis. Esta medida de la distancia es recomendada para el cálculo de valores atípicos en muestras multivariantes, ya que no sólo tiene en cuenta los valores de la observación a comparar sino su covarianza. Esta distancia se define con la siguiente ecuación:

$$D(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T * S^{-1} * (\vec{x}-\vec{y})}$$

Ecuación 11: cálculo de distancia Mahalanobis

   En esta ecuación $\vec{x}$ e $\vec{y}$ representan cada una de las dos observaciones a comparar de forma vectorial. La variable S^{-1} representa la matriz de covarianza invertida. Para calcular dicha matriz utilizaremos la herramienta que la fundación Apache proporciona para tal efecto (`org.apache.commons.math3.stat.correlation.Covariance`).

Aunque el resultado de esta operación sea un vector, tomaremos la distancia como la suma de cada uno de los valores del mismo.

3. Se ordenan las observaciones de menor a mayor distancia a la mediana (calculada en el paso anterior), para calcular los cuartiles Q1 y Q3, así como el rango intercuartílico (IQR). El cuartil 1 y el 3 son los resultantes de calcular la mediana de la primera y la segunda mitad de los datos respectivamente. El rango intercuartílico es calculado con la siguiente ecuación:

$$IQR = D(Q3) - D(Q1)$$

Ecuación 12: cálculo del rango intercuartílico

Nótese que para este cálculo sólo se tiene en cuenta la distancia calculada en el paso anterior, por lo que en la ecuación anterior tanto $D(x)$ representa la distancia a la mediana de la observación en la posición x.

4. Se eliminan de la base de datos todas aquellas observaciones que sean valores atípicos extremos, es decir, que no cumplan la siguiente ecuación:

$$D(Q1) - (3 * IQR) \leq \boldsymbol{D}(\boldsymbol{x}) \leq D(Q3) + (3 * IQR)$$

Ecuación 13: formula de detección de datos atípicos

En esta ecuación la variable x representa a cada una de las observaciones contenidas en la base de datos.

Se han eliminado de esta base de datos solamente las observaciones atípicas extremas dado que por la propia naturaleza de los datos, al no tener en algunas de las variables un límite superior por el que puedan ser acotados. Si eliminásemos también los valores atípicos leves, podríamos estar eliminando valores también importantes para el análisis.

3.5.1.3. Aislamiento de los datos a analizar

Debido a que Mahout no discrimina las variables del fichero de entrada, y analiza todos los datos como un conjunto, se deben proporcionar un fichero que únicamente contenga aquellos datos que queramos incluir en el proceso de agrupamiento. Por tanto, se debe llevar a cabo una eliminación de las columnas que no deban ser analizadas con anterioridad a la ejecución del algoritmo de agrupación. Esto es, dicho de otra manera, se deben aislar aquellas variables que deseamos analizar del resto de las variables, creando un nuevo fichero que sólo contenga aquellas columnas que sean de interés para el análisis que se vaya a realizar.

La elección de las columnas de interés dependerá del objetivo de cada uno de los análisis, y será explicado con posterioridad en el apartado 4, cuando describamos con exactitud los análisis que deseamos realizar sobre la base de datos de partes de accidentes.

El fichero resultante de esta limpieza estará en formato texto, conteniendo los datos de un siniestro en cada línea y separando las columnas por espacios. Es conveniente que el fichero esté en este formato para posteriormente ser convertido a un fichero secuencial utilizando las herramientas que proporciona para ello Mahout.

3.5.1.4. Normalización de los datos

Por el tamaño y la cantidad de los datos de nuestra muestra, con algunas de las variables sin límite superior como, por ejemplo, el coste de mano de obra de la reparación de un vehículo. De esta forma, podemos disponer de valores muy heterogéneos en este fichero de datos. Para que los datos sean comparables, será necesario que los normalicemos asegurándonos de que todos ellos estén en un rango, generalmente [0,1]. La fórmula que se utiliza para normalizar los datos es la siguiente:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} * 10^p$$

Ecuación 14: fórmula para normalización de datos

Dado que Mahout actúa con una precisión de 3 decimales a la hora de realizar el proceso de agrupamiento de los datos, el resultado de esta fórmula será multiplicado por una potencia de diez suficiente para dar la máxima precisión posible al proceso, modificando el intervalo especificado con anterioridad a $[0, 10^p]$. La elección del valor de p será justificada debidamente en cada parte del proceso de análisis de datos.

3.5.1.5. Carga del fichero en el sistema de ficheros distribuido

Una vez realizado todo el preprocesado manual, es el momento de copiar el fichero que contiene la base de datos al sistema de ficheros distribuido proporcionado por Hadoop. De esta forma podremos comenzar a utilizar las herramientas proporcionadas por Mahout para el tratamiento y análisis del mismo.

A partir de este momento, el fichero será accesible por todos los nodos que van a realizar el proceso de agrupamiento, de forma que cada uno de ellos pueda trabajar sobre estos datos de forma paralela.

3.5.1.6. Conversión a fichero secuencial

Para que el fichero se pueda procesar con Mahout, se deben utilizar las herramientas que él mismo proporciona para convertir el fichero de entrada, originalmente como vectores separados por espacios, a un fichero secuencial ordenado de forma aleatoria.

Esta conversión se lleva a cabo leyendo la base de datos línea a línea y escribiéndola con ayuda de la herramienta `org.apache.mahout.math.RandomAccessSparseVector` en el fichero de salida.

3.5.2. Agrupamiento de datos

El proceso de agrupamiento o clústering de los datos de los que disponemos se realizará en un proceso en dos fases como se explicó con anterioridad en el apartado 3.4.

Una vez se dispone de los datos correctamente preparados para realizar el proceso, en primer lugar se ejecutará el algoritmo *Canopy* incluido en Mahout, utilizando unos umbrales `U1` y `U2` adecuados para cada caso en concreto, por lo que serán elegidos y justificados debidamente en cada uno de los procesos de manera independiente. Este proceso creará una serie de clústeres, que como vimos son bastante imprecisos dado a que son tremendamente sensibles a los valores elegidos para los umbrales, como se comentó con anterioridad, pero el número de grupos creados por este algoritmo y el tamaño de cada uno de ellos nos permite saber cuál es el número de clústeres que se deben especificar al siguiente algoritmo.

Una vez decidido el número de grupos, simplemente falta ejecutar el algoritmo *k-Means* incluido en Mahout, proporcionándole como clústeres iniciales (necesarios para la ejecución de este algoritmo) los obtenidos con el algoritmo *Canopy*. El método de medida de distancia será elegido y justificado debidamente en cada uno de los estudios que se realicen sobre los datos, dado que la efectividad de cada uno de los métodos de medida depende exclusivamente de las características de las variables de la base de datos.

3.5.3. Postprocesado de datos de agrupamiento

Una vez ha terminado todo el proceso, Mahout nos indica en qué clúster situar cada uno de los vectores de la muestra que ha analizado. Sin embargo, nuestra base de datos tenía muchas otras variables que hemos omitido en el proceso, ya que no queríamos que fuesen consideradas para el proceso de agrupamiento.

Además, los datos que Mahout nos devuelve están normalizados y son los datos sin normalizar los que realmente resultan de interés para este estudio. Sin embargo, el proceso de normalización inverso puede ser impreciso dado que Mahout redondea los datos de entrada a tres posiciones decimales, lo cual es aceptable para el estudio, pero puede acarrear un error pequeño que se debe eliminar si es posible.

Por tanto, el proceso de tratamiento de los resultados de Mahout se basa en relacionar cada uno de los vectores que Mahout devuelve en los clúster de salida con la observación original contenida en el fichero CSV de entrada. Esto es fácilmente llevado a cabo dado que Mahout muestra el listado de vectores en la salida exactamente en el mismo orden en el que se encontraban en el fichero de entrada. De esta forma, sabremos qué observación original corresponde a cada clúster simplemente leyendo simultáneamente el archivo de salida de Mahout y el archivo de entrada

que se obtuvieron tras eliminar los datos atípicos, como se indicó en el apartado 3.5.1.2, ya que existirá una correspondencia entre el orden de las líneas de cada uno de ellos.

Una vez asociada cada observación con su grupo correspondiente, se creará un fichero en formato CSV que contenga todas las observaciones con cada una de sus variables por cada grupo que contenga únicamente las observaciones que le pertenezcan. Esto es equivalente a separar el fichero original en ficheros distintos por cada uno de los clúster.

Asimismo, se creará también un fichero que contenga los centroides de cada uno de los grupos, así como los radios de cada uno de ellos.

3.5.4. Creación de árboles de decisión

Para crear los árboles de decisión con Mahout se utilizará el algoritmo llamado *Random Forest* que se mencionó con anterioridad. Previamente a ello, es necesario preparar los ficheros de entrada, con el fin de crear dos ficheros en formato CSV separados por comas, sin línea de cabecera. El primero de estos ficheros servirá para realizar el entrenamiento, y el segundo para hacer pruebas.

El primer paso de la preparación de los datos es obtener los ficheros que se han extraído del postprocesado de datos del agrupamiento, eliminando la línea de cabeceras. Al igual que en el proceso de agrupamiento, en este caso es necesario eliminar todas las columnas que no se desea que sean tenidas en cuenta, dejando sólo las columnas que se deseen tener en cuenta para realizar la clasificación. Para continuar, se añadirá a cada uno de los ficheros una columna que identifica en qué grupo se encuentran. Una vez todos los ficheros están listos, se concatenarán todos ellos en un único fichero CSV.

Llegados a este punto es muy importante mezclar los resultados, de manera que el orden del fichero CSV sea aleatorio, previamente a realizar la separación en dos bases de datos (de entrenamiento y de pruebas), ya que se debe garantizar que existan observaciones de todos los clúster dentro de ambas bases de datos. La separación en dos bases de datos es realizada con ayuda del comando `split` de Mahout. Se ha decidido que en todos los casos se tomarán el 70% de los datos para entrenar el proceso de clasificación y el 30% para realizar pruebas.

El siguiente paso en este proceso consiste en copiar las dos bases de datos al sistema de ficheros distribuido de Hadoop, para permitir continuar con los siguientes pasos del análisis, que ya pueden ser realizados de forma distribuida.

Para continuar, se debe crear un fichero que contenga los metadatos de la base de datos que va a ser analizada. Para ello Mahout incorpora una herramienta llamada `Describe` en el paquete `org.apache.mahout.classifier.df.tools`, a la cual se debe proporcionar el tipo de cada una de las variables de entre tres opciones: numérica, texto o etiqueta. La última de estas tres (etiqueta)

es utilizada para identificar la columna que contiene la identificación del clúster, es decir, la columna que indica a qué clúster pertenece cada observación.

Una vez realizado el fichero con los metadatos, se puede proceder a realizar el entrenamiento del proceso de clasificación `BuildForest`. A este proceso hay que proporcionarle la base de datos que contiene los datos de entrenamiento, así como el fichero que contiene los metadatos de la misma. Cuando el proceso de creación de los árboles termina, Mahout genera un fichero llamado `forest.seq`, el cual servirá para clasificar nuevas observaciones de manera automática.

Por último, se debe probar los árboles generados con aquellas observaciones que han sido reservadas para hacer pruebas. Esto se realiza con la herramienta `TestForest`, la cual está incluida en el paquete `org.apache.mahout.classifier.df.mapreduce`. Este proceso mostrará las estadísticas sobre las instancias clasificadas correctamente, las instancias clasificadas de manera incorrecta, así como la matriz de confusión. Adicionalmente, también presenta cierta información sobre dicha matriz: precisión (porcentaje de observaciones clasificadas correctamente, valor *Kappa* (precisión observada frente a precisión esperada), fiabilidad del bosque y desviación típica de dicha fiabilidad.

Con toda esta información, se dispondrá de información suficiente para comprobar si los árboles generados son fiables, y si esto es así, para clasificar nuevas observaciones que se vayan incluyendo a la base de datos.

4. Resultados obtenidos

En este apartado se llevará a la práctica todo aquello que ha sido planteado con anterioridad, de manera que se obtengan todas las conclusiones necesarias para este estudio, las cuales fueron definidas en el apartado introductorio.

4.1. Creación de grupos de severidad

Desde el punto de vista de una aseguradora, es interesante poder categorizar los distintos partes de siniestro según su coste de mano de obra, pintura y sustitución de piezas y posteriormente analizar los resultados de manera exhaustiva para poder entender qué características comunes tienen entre sí los siniestros de cada uno de estos grupos.

Si analizamos los datos aislando estas tres variables, podremos ver qué tipos de siniestros existen en función del coste de los mismos. No se puede saber a priori qué tipo de conclusiones se pueden extraer de este análisis, pero es bastante lógico que aparezcan grupos de costes de reparación de siniestros que tendrán características similares entre ellos.

Adicionalmente, se puede profundizar en dicho análisis si dentro de cada grupo resultante tratamos de agrupar los datos según las zonas de impacto, pues sería interesante ver si estos grupos tienden a tener un patrón de zonas de impacto común o si esto no es así.

4.1.1. Proceso de obtención de los grupos de severidad

En un primer lugar, se han obtenido los distintos grupos de severidad de los siniestros, aislando las tres variables `Tot_mo`, `Tot_pint` y `Tot_sust` (coste de mano de obra, coste de pintura y coste de sustitución de piezas respectivamente). A partir de estos datos se ha realizado el preprocesado indicado en el apartado 3.5.1. Con este preprocesado se han eliminado un total de 19.992 valores atípicos (6.08% de los datos), lo cual está dentro de los rangos recomendados habitualmente (entre 0 y 10% de los datos).

Previamente al procesado, es imprescindible realizar ciertos análisis estadísticos sobre los datos, que nos sirven para hacernos una idea de su naturaleza, y poder elegir mejor los parámetros para el proceso de agrupamiento. Los resultados estadísticos obtenidos sobre las variables sin normalizar se pueden observar en la tabla 5.

	Tot_mo *Coste de mano de obra*	Tot_pint *Coste de pintura de piezas*	Tot_sust *Coste de piezas sustituidas*
Mínimo	0,00 €	0,00 €	0,00 €
Máximo	24.800,00 €	1.091,80 €	3.672,30 €
Media	176,51 €	246,62 €	261,68 €
Cuartil 1	58,28 €	104,40 €	8,97 €
Mediana	111,00 €	198,51 €	91,59 €
Cuartil 3	228,00 €	319,38 €	279,99 €
Desv. Típica	186,62 €	217,79 €	455,15 €

Tabla 5: análisis sobre los costes de los siniestros

Nótese que los datos han sido normalizados utilizando el rango [0, 10.000] para aportar precisión al experimento, dado que el máximo valor que encontramos en la base de datos es 24.800,00 y con este intervalo aportaremos suficiente precisión para tener en cuenta todas las posiciones decimales de la base de datos original, ya que Mahout tendrá en cuenta las tres primeras posiciones decimales de cada variable.

En estos datos se puede observar alguna información que caracteriza los datos que vamos a analizar. En primer lugar, es notorio que el rango en el que se mueve el coste de mano de obra es el 0-24.800 euros, mientras el coste de pintura y sustitución oscilan un rango mucho más reducido (0-1.091,80 y 0-3.672,30 respectivamente). A pesar de que a priori se puede pensar que los valores del coste de pintura son más altos, podemos ver que la media del mismo es inferior a la de los costes de pintura y sustitución. Esto, en conjunto con el valor de la mediana, nos permite comprender que el coste de mano de obra de esta base de datos es, en la mayoría de los casos, muy inferior a este valor.

En esta tabla también se puede observar que en los tres casos la mediana es inferior a la media, lo cual nos indica que hay más costes bajos que altos en la muestra, y los valores altos están más dispersos. De hecho, el tercer cuartil en todos ellos es bastante cercano a dicho valor, lo que refuerza esta afirmación.

También podemos observar que la desviación típica de las tres variables no es muy alta, dados los rangos en los que oscilan las mismas. De todas formas, la desviación típica de los costes de sustitución de piezas dobla la de los otros dos costes, lo cual significa que es una variable más dispersa por lo general.

A continuación se ha realizado un análisis de correlaciones entre las variables que se están estudiando, el cual nos indicará cómo están relacionadas entre sí estas variables.

	Tot_mo *Coste de mano de obra*	Tot_pint *Coste de pintura de piezas*	Tot_sust *Coste de piezas sustituidas*
Tot_mo *Coste de mano de obra*	1	0,5483	0,5661
Tot_pint *Coste de pintura de piezas*	0,5483	1	0,1931
Tot_sust *Coste de piezas sustituidas*	0,5661	0,1931	1

Tabla 6: correlación entre los costes de los siniestros

En esta tabla se puede observar cómo la correlación entre todas estas variables es positiva, es decir, generalmente un incremento de una de las tres variables supondrá también una subida en las demás. Esto demuestra que existe cierta relación lineal entre estas variables, de diferentes magnitudes. Los costes de pintura están correlacionados de una forma moderada con los demás gastos, de manera que su relación lineal es similar y de una magnitud media. Sin embargo, la correlación entre los costes de pintura y de sustitución es muy baja.

Si representamos gráficamente la dispersión de las variables podremos observar claramente lo indicado anteriormente, así como otros detalles que no se aprecian sólo con los datos estadísticos. A continuación observaremos la representación de la dispersión de estas tres variables, utilizando el eje horizontal únicamente para crear una segunda dimensión, distribuyendo uniformemente a lo largo de dicho eje los datos de la muestra, lo que nos permitirá ver los datos con una mayor claridad.

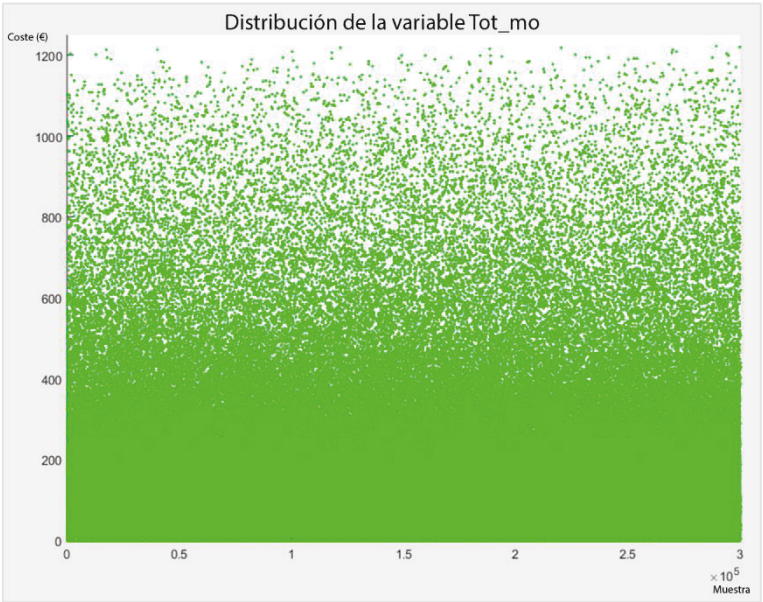


Ilustración 18: dispersión de la variable Tot_mo

En el caso de la variable `Tot_mo` es importante aclarar que se muestra sólo el rango [0-1.250] porque las apariciones superiores a este valor son muy dispersas y si son representadas impiden apreciar la distribución de los valores más comunes. Si se observa detenidamente esta gráfica se

puede ver como los costes más frecuentes son los más bajos, y cuanto más altos son menos frecuentes.

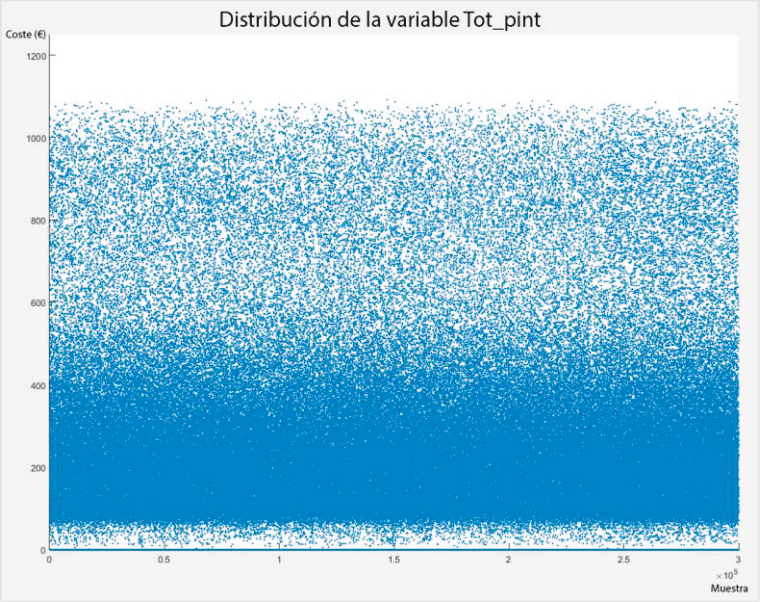


Ilustración 19: dispersión de la variable Tot_pint

En esta gráfica se puede observar claramente como en el coste de pintura el intervalo [50, 300] aproximadamente es el más frecuente, mientras que los valores por debajo y por encima de dicho rango son más dispersos cuanto más alejados se encuentran de este rango. Por otra parte, se puede apreciar que existe una franja de datos que se encuentran en el valor 0 (siniestros sin coste de mano de obra), mientras que en el rango (0-50] el número de siniestros es muy bajo, lo que demostraría que, por lo general, si existen gastos de mano de obra, éstos no tendrán un coste inferior a 50 euros.

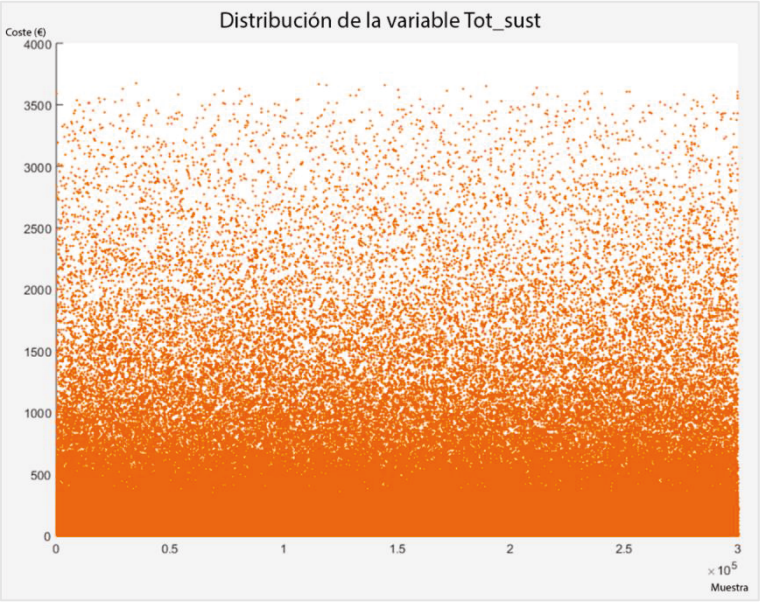


Ilustración 20: dispersión de la variable Tot_pint

En el caso de los costes de pintura se puede observar un comportamiento muy similar al de los costes de mano de obra, con la diferencia de que estos datos oscilan en un rango tres veces más grande. Igual que en el caso anterior, los costes bajos son los más frecuentes y se van haciendo menos frecuentes a medida que los costes se elevan.

Habiendo estudiado los datos de entrada correctamente, es el momento de ejecutar el algoritmo de agrupamiento *Canopy*, que permite estimar el número de grupos en los que se puede separar la muestra, proporcionando al mismo los siguientes parámetros:

- *Método de medida de distancia*: se ha decidido emplear la distancia Euclídea, ya que al considerar la distancia como la longitud de la línea recta que une los puntos a estudiar, es apropiado utilizarla debido a que se está tratando un espacio tridimensional en el cual las tres variables son de la misma naturaleza, y la distancia en línea recta es la distancia real en euros entre ambos vectores.
- *Umbral 1*: se ha establecido como el valor 3.500,00. De esta forma, serán considerados para formar parte de un clúster todos aquellos valores cuya distancia al centroide del mismo sea menor que 3.500 euros. Este valor ha sido elegido porque es una distancia elevada, y así el proceso *Canopy* podrá considerar un vector como perteneciente a un clúster aunque su distancia sea bastante elevada. De todas formas, hay que recordar que el cumplimiento de la restricción del primer umbral no es definitiva, los vectores pueden ser movidos si se encuentra que tienen una menor distancia con otro centroide.
- *Umbral 2*: se ha establecido como 100,00 euros. De esta forma, todas aquellas observaciones cuya distancia al centroide de un grupo sea inferior a este valor serán consideradas como pertenecientes a un clúster y no serán movidas. Se ha elegido este valor porque es un valor pequeño y razonable, con el que podríamos decir que dos siniestros cuyos costes disten 100 euros o menos tienen un precio similar.

La ejecución de este algoritmo, la cual ha tardado un total de 12,48 minutos, nos muestra que existen 4 clústeres claramente definidos, mientras que uno de todos los valores no encaja en ninguno de los grupos. Una vez obtenida esta información, el siguiente paso es ejecutar sobre estos datos el algoritmo *k-Means* utilizando como clústeres de partida los obtenidos por el algoritmo *Canopy*, el mismo método de medida de distancia que anteriormente y estableciendo que deseamos obtener un total de 4 grupos. La ejecución de este proceso tuvo una duración de 8,10 minutos.

Una vez se dispone de los distintos grupos de severidad de los siniestros de nuestra base de datos, puede ser interesante analizar qué tipo de impacto corresponde a cada uno de los grupos, para analizar si existe alguna relación entre los grupos de severidad y las zonas donde se han recibido los impactos.

De este análisis se desean obtener resultados que nos indiquen qué caracteriza a cada uno de estos grupos de sinestros. Por ejemplo, se puede hipotéticamente encontrar un grupo de siniestros de severidad baja que presumiblemente esté basado en raspones en zonas muy superficiales de los vehículos, y no afecten a piezas de mayor valor económico. Este tipo de hipótesis tienen un razonamiento lógico, pero con el análisis de los datos de que se dispone se pretende convertir estas hipótesis en hechos probados. También, se pretenden encontrar características desconocidas, pero que resulten interesantes para el análisis de los siniestros de estos grupos.

En este caso, se realizarán cuatro análisis distintos, uno por cada grupo de severidad obtenido. En cada uno de estos análisis, se proporcionarán a Mahout las 710 variables binarias que indican cuáles de las piezas han sido reparadas, cuáles pintadas y cuáles sustituidas.

Al igual que en el caso anterior, se realizará el preprocesado de datos, pero no será necesario realizar ningún tipo de normalización a los mismos ya que el proceso no realizaría ningún cambio en los datos, al tener todos valores 0 o 1.

Para continuar, como se ha explicado con anterioridad, se ejecutará el algoritmo *Canopy* sobre estas variables. En este caso, se han elegido los siguientes parámetros para los cuatro grupos de severidad:

- *Método de medida de distancia*: se ha decidido emplear la distancia *Manhattan*, la cual es idónea para datos binarios de este tipo, ya que medirá la distancia entre dos vectores dando como resultado un número que equivale al número de variables que son diferentes entre los dos vectores.
- *Umbral 1*: se ha establecido como el valor 50. De esta forma, serán considerados para formar parte de un clúster todos aquellos valores cuya distancia al centroide del mismo sea menor que este número, es decir, que haya 50 o menos piezas afectadas de diferencia. Este número permitirá al proceso *Canopy* considerar para formar parte de un clúster a vectores cuya distancia sea bastante grande sin comprometer el tiempo de respuesta del programa.
- *Umbral 2*: se ha establecido como 5. De esta forma, todas aquellas observaciones cuya distancia al centroide de un grupo sea inferior a este valor serán consideradas como pertenecientes a un clúster y no serán movidas, ya que se considera que una diferencia de 5 piezas afectadas entre los distintos siniestros que se hallen en un grupo algo muy pequeño.

En este caso, dado que tenemos un número muy grande de variables, y el algoritmo *Canopy* es bastante sensible a los umbrales elegidos, se utilizará el número de clústeres que resulten como

aproximado, y se repetirá el método *k-Means* en diversas ocasiones, variando el número de grupos a obtener para afinar el resultado. De esta forma, si al variar el número de grupos se observa que uno de los grupos da resultados con poca relación entre ellos, o con una cantidad muy pequeña de observaciones en su interior, sabremos que dicho número de grupos no es válido.

4.1.2. Análisis de los grupos de severidad obtenidos

En la siguiente ilustración se ha representado gráficamente la distribución de observaciones que se obtiene aplicando este método de agrupamiento de datos.

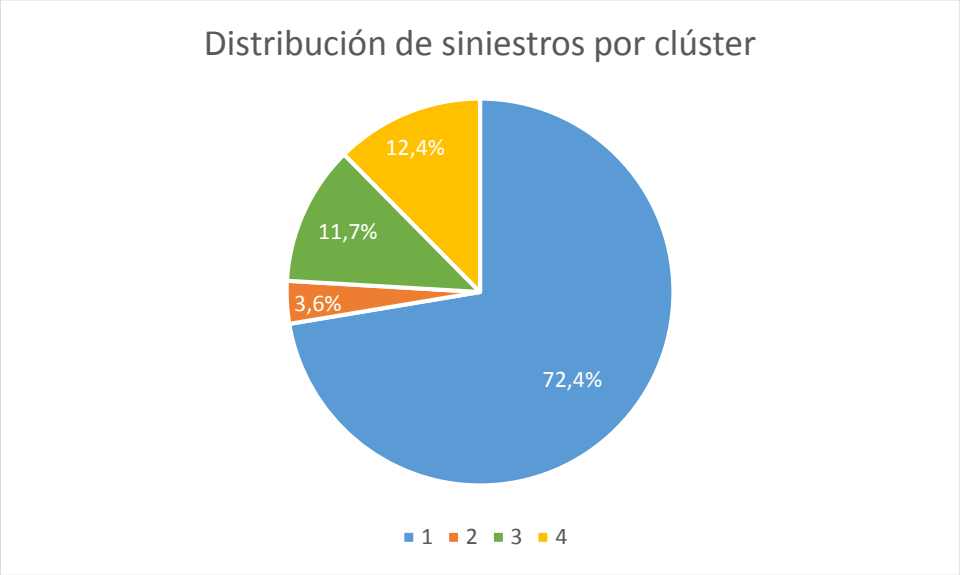


Ilustración 21: distribución de siniestros por severidad

Como se puede observar antes de ahondar en los detalles de cada uno de los cuatro clúster resultantes, existe un gran grupo de siniestros que contiene casi tres cuartas partes de las observaciones, mientras que el resto de los siniestros se distribuyen en otros tres grupos de un tamaño inferior (12,4%, 11,7% y 3,6%) lo cual nos demuestra que la mayor parte de los siniestros siguen un patrón común. Sin embargo, será necesario profundizar en el análisis para comprender cuales son las características particulares de cada uno de estos grupos.

Adicionalmente, se ha realizado una representación en tres dimensiones de la distribución de los costes generados por los partes de accidente en función del clúster al que pertenece cada uno.

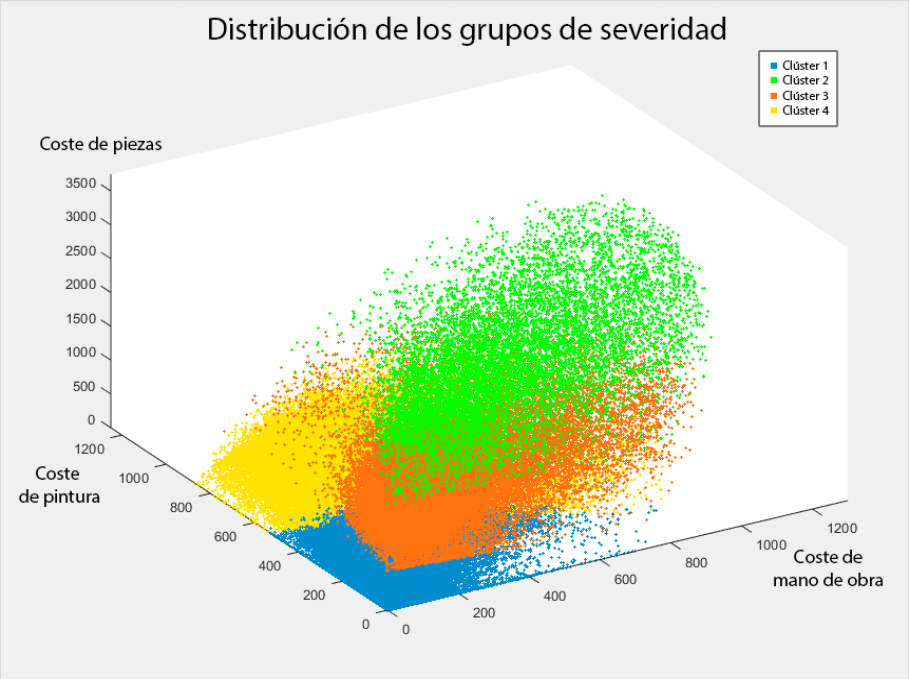


Ilustración 22: distribución de costes de siniestros por severidad

4.1.2.1. Grupo 1: severidad baja

Se han obtenido ciertas estadísticas que representan los datos contenidos en este clúster, como se puede observar en la siguiente tabla.

	Tot_mo *Coste de mano de obra*	Tot_pint *Coste de pintura de piezas*	Tot_sust *Coste de piezas sustituidas*
Mínimo	0,00 €	0,00 €	0,00 €
Máximo	827,33 €	526,77 €	582,63 €
Media	99,69 €	157,58 €	102,71 €
Cuartil 1	47,98 €	87,24 €	4,65 €
Mediana	78,43 €	156,62 €	35,60 €
Cuartil 3	131,97 €	231,67 €	189,58 €
Desv. Típica	76,21 €	107,52 €	121,78 €

Tabla 7: distribución de costes en siniestros de severidad baja

Lo primero que salta a la vista al observar estos resultados es que el primer clúster contiene los máximos, mínimos, medias, medianas y cuartiles más pequeños, comparado con el resto de los grupos. Esto, junto a la representación gráfica del mismo, nos demuestra, en conjunto con los datos de la ilustración 22, que la mayor parte de los siniestros ocurridos a este vehículo han sido de una severidad baja. Además, este clúster muestra una desviación típica bastante bajas comparada con el resto, lo cual nos demuestra que los costes de este tipo de siniestros no son muy variables. En este grupo podríamos encontrar, por ejemplo, siniestros ocasionados por raspones o choques leves, donde hay que hacer una serie de pequeñas reparaciones en pocas

piezas, y sustituciones de piezas pequeñas o superficiales. Por ello, denominaremos este grupo como *Severidad baja*.

A continuación se ha realizado un análisis de correlaciones entre las variables que se están estudiando en este clúster, el cual nos indicará cómo están relacionadas entre sí estas variables.

	Tot_mo *Coste de mano de obra*	Tot_pint *Coste de pintura de piezas*	Tot_sust *Coste de piezas sustituidas*
Tot_mo *Coste de mano de obra*	1	0,5172	0,1057
Tot_pint *Coste de pintura de piezas*	0,5172	1	-0,1443
Tot_sust *Coste de piezas sustituidas*	0,1057	-0,1443	1

Tabla 8: correlación entre los costes en siniestros de severidad baja

En esta tabla se puede observar cómo se ha reducido correlación entre todas estas variables con respecto a la base de datos completa. Se puede ver cómo sigue existiendo una correlación positiva y moderada entre el coste de pintura y el coste de mano de obra, lo que indica que estas variables están bastante influenciadas: el incremento de costes de mano de obra suele implicar un incremento en costes de pintura y viceversa. Adicionalmente, se puede observar una correlación muy baja entre los costes de sustitución de piezas y los demás costes, lo cual indica que estas variables no son muy dependientes entre sí. En concreto se puede observar que los costes de mano de obra y los costes de sustitución de piezas están correlacionados positivamente, de manera que la relación es lineal (si unos costes suben, los otros también). Sin embargo, los costes de pintura y los de sustitución de piezas están correlacionados negativamente (si uno sube, el otro baja).

Tras la ejecución de los procesos de agrupamiento descritos con anterioridad, se ha determinado que existen 5 tipos de siniestros claramente definidos en siniestros de severidad baja, que se distribuyen como se puede observar en la siguiente ilustración.

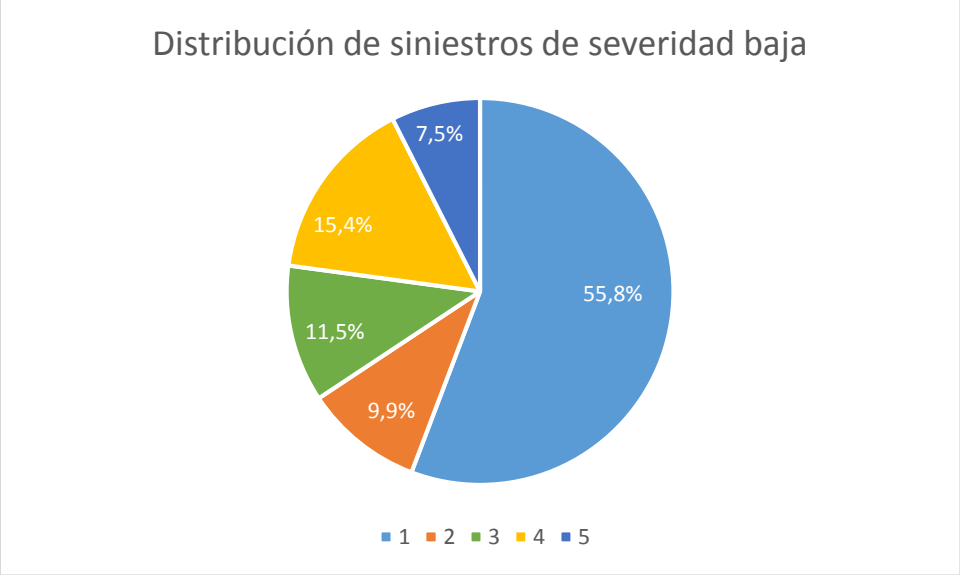


Ilustración 23: distribución de los siniestros de severidad baja

Se puede observar como en este grupo de siniestros existe un grupo claramente predominante, que contiene algo más de la mitad de los siniestros, y otros 4 grupos que contienen menor cantidad de datos.

Para continuar se ha elaborado una tabla que define estos 5 clústeres que han sido creados, especificando las 10 órdenes de reparación, sustitución o pintura que son realizadas con mayor frecuencia en este tipo de siniestros.

Clúster	Siniestros	Variables significativas		Porcentaje
1: Golpe frontal	55,77%	PT_63203	Pintura de paragolpes delantero	35,42%
		REP_63203	Reparación de paragolpes delantero	18,14%
		PT_63303	Pintura de paragolpes trasero	14,79%
		SUST_64103	Sustitución de parabrisas delantero	14,36%
		SUST_63203	Sustitución de paragolpes delantero	13,22%
		PT_50101	Pintura de aleta delantera izquierda	10,50%
		SUST_66953	Sustitución anagrama del fabricante	10,49%
		PT_55103	Pintura de capó	10,15%
		PT_50102	Pintura de aleta delantera derecha	8,53%
		REP_63303	Reparación de paragolpes trasero	8,53%

Tabla 9: zonas de impacto en siniestros de severidad baja (I)

Clúster	Siniestros	Variables significativas		Porcentaje
2: Golpe lateral izquierdo	9,92%	PT_57101	Pintura puerta delantera izquierda	97,42%
		SUST_66201	Sustitución molduras izquierda	85,85%
		REP_57101	Reparación puerta delantera izquierda	74,86%
		PT_50101	Pintura aleta delantera izquierda	65,80%
		PT_53101	Pintura aleta trasera izquierda	43,71%
		PT_58101	Pintura de puerta trasera izquierda	36,24%
		REP_50101	Reparación aleta delantera izquierda	35,24%
		REP_53101	Reparación aleta trasera izquierda	30,29%
		REP_58101	Reparación puerta trasera izquierda	26,05%
		PT_66501	Pintura retrovisor izquierdo	17,22%
3: Golpe lateral derecho	11,46%	SUST_66202	Sustitución de molduras derecha	88,77%
		PT_57102	Pintura de puerta delantera derecha	83,45%
		PT_53102	Pintura de aleta trasera derecha	68,22%
		REP_57102	Reparación puerta delantera derecha	65,71%
		REP_53102	Reparación de aleta trasera derecha	57,91%
		PT_50102	Reparación de aleta delantera derecha	53,29%
		PT_58102	Pintura de puerta trasera derecha	52,95%
		REP_58102	Reparación de puerta trasera derecha	44,31%
		REP_50102	Reparación aleta delantera derecha	29,71%
		PT_63303	Pintura de paragolpes trasero	21,49%
4: Golpe trasero	15,39%	PT_63303	Pintura de paragolpes trasero	99,71%
		SUST_63303	Sustitución de paragolpes trasero	68,60%
		REP_63303	Reparación de paragolpes trasero	45,47%
		PT_55303	Pintura de portón trasero	33,19%
		SUST_66953	Sustitución anagrama del fabricante	31,47%
		SUST_66303	Sustitución de rejilla del radiador	30,85%
		REP_55303	Reparación de portón trasero	23,09%
		SUST_94201	Sustitución de faro trasero izquierdo	21,28%
		PT_53203	Pintura de faldón trasero	19,95%
		REP_53203	Reparación de faldón trasero	19,69%
5: Golpe trasero-izquierdo	7,47%	PT_53101	Pintura de aleta trasera izquierda	96,38%
		REP_53101	Reparación de aleta trasera izquierda	82,73%
		PT_63303	Pintura de paragolpes trasero	71,32%
		SUST_66201	Sustitución de molduras izquierda	46,48%
		PT_58101	Pintura de puerta trasera izquierda	31,94%
		REP_63303	Reparación de paragolpes trasero	29,86%
		REP_58101	Reparación puerta trasera izquierda	22,56%
		SUST_63303	Sustitución de paragolpes trasero	18,21%
		PT_57101	Pintura de puerta delantera izquierda	15,37%
		SUST_63301	Sustitución paragolpes trasero izquierdo	14,33%

Tabla 10: zonas de impacto en siniestros de severidad baja (II)

En la siguiente ilustración se pueden observar los 5 grupos que se han obtenido dibujados sobre un vehículo, de forma que se aprecia de una manera gráfica lo que representa la tabla anterior.

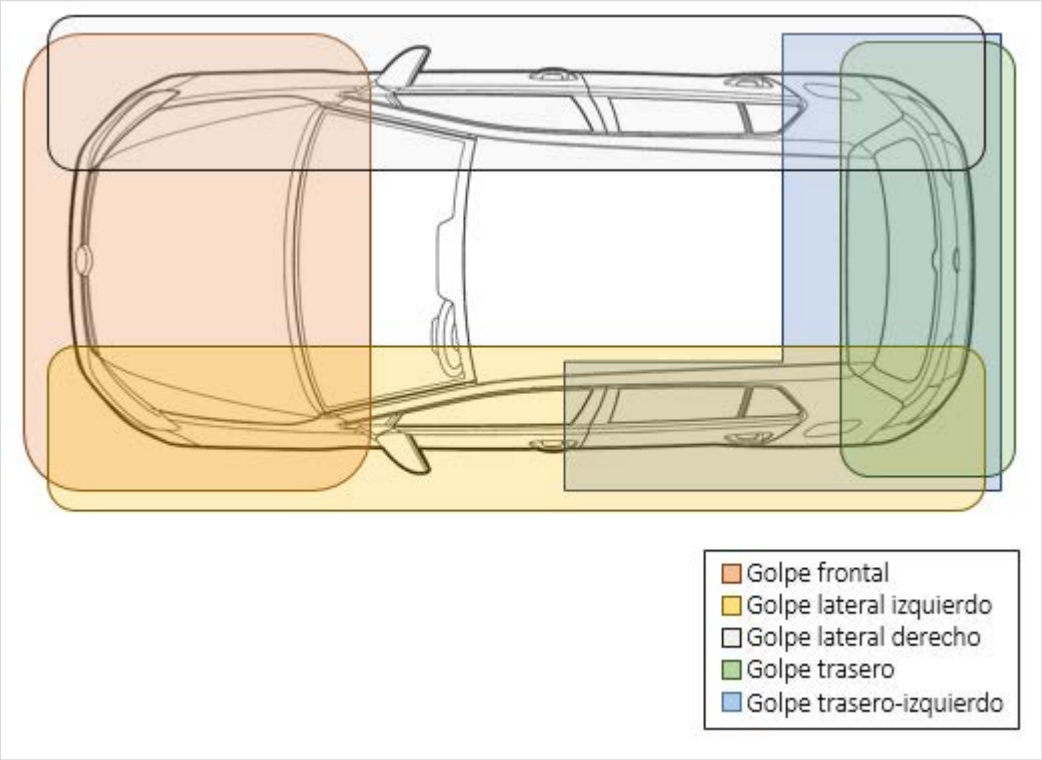


Ilustración 24: zonas de impacto en siniestros de severidad baja

La primera conclusión que se puede extraer de estos datos es que los golpes frontales son los que están peor definidos, es decir, suelen verse implicadas piezas distintas, y no existe un gran predominio de una o varias piezas afectadas por el impacto. Esto puede estar originado principalmente porque existe un gran número de observaciones en este grupo, y por tanto existir mayor variedad. Sin embargo, es importante tener en cuenta que los turismos (categoría del vehículo que está siendo estudiado), por lo general tienen más piezas y de mayor valor en la parte frontal. Esto implica que al producirse un golpe frontal hay más piezas con riesgo de ser dañadas que, por ejemplo, con un choque lateral.

En el caso de los golpes frontales se observa como es habitual la necesidad de pintar o reparar el paragolpes delantero, lo cual es debido seguramente a que está diseñado para absorber los golpes y proteger las piezas que lo rodean. También se puede observar que aparece con cierta frecuencia el paragolpes trasero, lo cual puede significar que el golpe ha sido a la vez por la parte delantera y la parte trasera del vehículo, por ejemplo con un choque en cadena.

El siguiente golpe más común es el golpe trasero, que en conjunto con los golpes trasero-izquierdos albergan aproximadamente el 23% de los siniestros de severidad baja. En el caso del golpe trasero, los daños en el paragolpes trasero son casi imprescindibles (99.7% de los vehículos requieren pintura del mismo). Esto es lógico porque el paragolpes siempre es el elemento que más sobresale del vehículo por la parte trasera, precisamente para absorber los posibles golpes que reciba el vehículo. Por otra parte, el portón del maletero también implica gastos de pintura

en algunas ocasiones, las cuales son generalmente por golpes más fuertes, en los que la defensa no sea capaz de absorber el golpe completo, o con un vehículo u obstáculo que se encuentre a distinta altura de la defensa.

En el caso del golpe trasero-izquierdo se incluyen algunas zonas del lateral izquierdo del vehículo, como la aleta o la puerta trasera izquierda. Todas ellas son pintadas o reparadas, lo cual significa que los golpes no han sido muy fuertes. En el caso de las molduras, son piezas pequeñas que no suelen ser reparadas ya que es más costoso hacerlo que colocar unas nuevas.

En el caso de los golpes laterales, se ve que es ligeramente más frecuente el golpe lateral derecho, quizás debido a que el conductor va sentado en el lado izquierdo del vehículo y tiene un mayor control sobre lo que ocurre en el lateral en el que circula. En el caso del lateral izquierdo, predominan los daños en la puerta delantera y las molduras delantera y trasera. En el caso del golpe lateral derecho, las piezas afectadas suelen ser las molduras, las puertas y la aleta trasera. En ambos casos, las piezas afectadas son de poco valor (por ejemplo molduras) o son sustituidas o pintadas, lo que significa que no se trata de impactos muy fuertes.

En los cinco casos, se observa claramente como las piezas más comúnmente afectadas son piezas muy superficiales, y generalmente los trabajos son de reparación o de pintura en piezas grandes (paragolpes, aletas, portones…), o su sustitución en caso de piezas pequeñas (molduras), lo cual nos indica que se trata de siniestros pequeños, en los que el vehículo aparentemente ha sufrido daños superficiales, que seguramente estén ocasionados por raspones o bien con otros vehículos o bien con columnas, muros, quitamiedos, etcétera.

4.1.2.2. Grupo 2: severidad alta

Se han obtenido ciertas estadísticas que representan los datos contenidos en este clúster, como se puede observar en la siguiente tabla.

	Tot_mo *Coste de mano de obra*	Tot_pint *Coste de pintura de piezas*	Tot_sust *Coste de piezas sustituidas*
Mínimo	0,00 €	0,00 €	0,00 €
Máximo	24.800,00 €	1.069,50 €	3.672,30 €
Media	583,20 €	387,62 €	2.083,40 €
Cuartil 1	376,72 €	288,22 €	1.658,80 €
Mediana	544,88 €	385,34 €	1.946,40 €
Cuartil 3	764,97 €	490,57 €	2.400,70 €
Desv. Típica	347,13 €	181,50 €	520,04 €

Tabla 11: distribución de costes en siniestros de severidad alta

En el caso del segundo clúster, podemos ver como los valores que toman los costes de los siniestros son los más elevados, por lo general, de todos los siniestros. En concreto, el coste que

es característicamente alto, como se puede observar gráficamente, es el coste de sustitución de piezas en el vehículo. Estos datos se ven reforzados por las estadísticas que se han obtenido con anterioridad, donde se puede observar que tanto la media, como los cuartiles y la mediana de los datos son más elevadas que el resto, excepto en el caso de los costes de pintura, donde es más bajo. Con respecto a la desviación típica de los datos, se puede apreciar que, exceptuando también los costes de pintura, son los más elevados de todos los clúster. En este caso, un siniestro de este tipo suele implicar un golpe fuerte de difícil reparación, como un accidente a gran velocidad en el que varias piezas se ven dañadas y necesitan ser sustituidas y/o pintadas. Por todo esto, se considerará a este clúster como *Severidad alta*.

A continuación se ha realizado un análisis de correlaciones entre las variables que se están estudiando en este clúster, el cual nos indicará cómo están relacionadas entre sí estas variables.

	Tot_mo *Coste de mano de obra*	Tot_pint *Coste de pintura de piezas*	Tot_sust *Coste de piezas sustituidas*
Tot_mo *Coste de mano de obra*	1	0,4119	0,1161
Tot_pint *Coste de pintura de piezas*	0,4119	1	0,0167
Tot_sust *Coste de piezas sustituidas*	0,1161	0,0167	1

Tabla 12: correlación entre los costes en siniestros de severidad alta

En esta tabla se puede observar cómo se ha reducido en gran medida la correlación entre todas estas variables con respecto a la base de datos completa. Esto tiene sentido dado que se ve que es un grupo muy disperso. Se puede ver cómo sigue existiendo una correlación positiva y moderada entre el coste de pintura y el coste de mano de obra, lo que indica que estas variables están bastante influenciadas: el incremento de costes de mano de obra suele implicar un incremento en costes de pintura y viceversa. Adicionalmente, se puede observar una correlación muy baja entre los costes de sustitución de piezas y los demás costes, lo cual indica que estas variables no son muy dependientes entre sí. En concreto se puede observar que los costes de mano de obra y los costes de sustitución de piezas están correlacionados de forma muy baja y positiva, de manera que la relación es lineal (si unos costes suben, los otros también pero en una medida muy reducida). Sin embargo, los costes de pintura y los de sustitución de piezas casi no están correlacionados, por lo que se podría decir que no existe correlación lineal, aunque no se sabe si existirá correlación de otro tipo.

En este grupo de siniestros se han podido formar un total de 7 subgrupos de zonas de impacto, en los cuales existen diversos tipos de siniestros muy bien definidos. Este gran número de grupos era de esperar dado que el grupo de siniestros de severidad alta es el más disperso de todos los

grupos de severidad que se han creado. En la siguiente ilustración se puede observar la distribución de zonas de impacto en este tipo de siniestros.

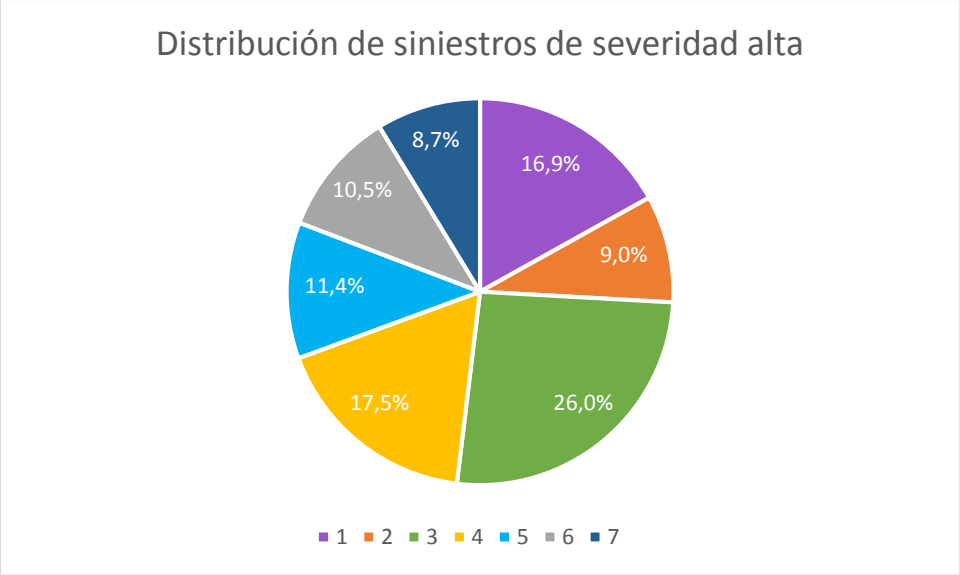


Ilustración 25: distribución de zonas de impacto en siniestros de severidad alta

A continuación se detallarán uno a uno cada uno de estos grupos de zonas de impacto en siniestros de severidad alta, teniendo en cuenta las 10 órdenes de pintura, sustitución o reparación que se muestran con más frecuencia en estos grupos. Sin embargo, sería un error tener en cuenta sólo las 10 primeras órdenes en grupos en los que haya órdenes que aparezcan en al menos el 70% de los siniestros. Por ello, se incluirán todas las órdenes con un porcentaje de aparición que superen dicho porcentaje, aunque se superen las 10 órdenes mencionadas con anterioridad.

Grupo 1: golpe leve con alto número de piezas implicadas (vehículos de 5 puertas)

	Variables significativas	Porcentaje
PT_58102	Pintura de puerta trasera derecha	98,89%
PT_58101	Pintura de puerta trasera izquierda	97,95%
PT_53102	Pintura de aleta trasera derecha	95,27%
PT_53101	Pintura de aleta trasera izquierda	95,13%
PT_63303	Pintura de paragolpes trasero	92,60%
PT_57102	Pintura de puerta delantera derecha	90,13%
PT_63203	Pintura de paragolpes delantero	88,70%
PT_55103	Pintura de capó	88,17%
PT_50101	Pintura de aleta delantera izquierda	87,08%
PT_55303	Pintura de portón trasero	84,57%
PT_57101	Pintura de puerta delantera izquierda	84,43%
PT_50102	Pintura de aleta delantera derecha	82,04%
SUST_66303	Sustitución de rejilla del radiador	76,95%
SUST_66202	Sustitución de molduras derecha	73,74%
SUST_66201	Sustitución de molduras izquierda	72,40%

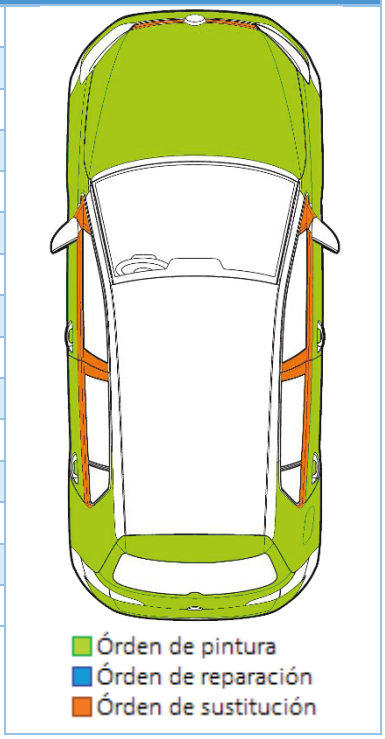


Tabla 13: zona de impacto en siniestros de severidad alta (grupo 1)

En este caso se puede observar con claridad que en su totalidad se trata de órdenes de pintura en todas las piezas que rodean el vehículo, incluyendo puertas, aletas, paragolpes, capó y el portón trasero. Además, se observan también varias órdenes de sustitución de piezas de pequeño valor, donde el coste de sustitución es inferior al de reparación o pintura de las piezas, o en algunos casos simplemente la reparación o pintura de las mismas no es posible.

Este tipo de siniestro se puede deber a rozaduras leves con otros vehículos, golpes leves ocasionados por pérdida del control del vehículo, como por aquaplaning, etcétera. En este caso los daños también pueden ser debidos al vandalismo, a condiciones meteorológicas adversas o a la circulación por carreteras no habilitadas para tal efecto. Si se observa el coste de pintura de estas piezas individualmente, se puede apreciar que no tiene un coste muy elevado, pero dado el elevado número de piezas afectadas, el coste total de la reparación sí que es elevado.

Grupo 2: golpe leve con alto número de piezas implicadas (vehículos de 3 puertas)

Variables significativas		Porcentaje
PT_55103	Pintura de capó	89,40%
PT_57102	Pintura de puerta delantera derecha	87,77%
PT_50101	Pintura de aleta delantera izquierda	85,41%
PT_53102	Pintura de aleta trasera derecha	83,76%
PT_53101	Pintura de aleta trasera izquierda	82,85%
PT_57101	Pintura de puerta delantera izquierda	82,59%
PT_63303	Pintura de paragolpes trasero	82,07%
PT_50102	Pintura de aleta delantera derecha	81,72%
PT_63203	Pintura de paragolpes delantero	81,43%
PT_55303	Pintura de portón trasero	80,96%

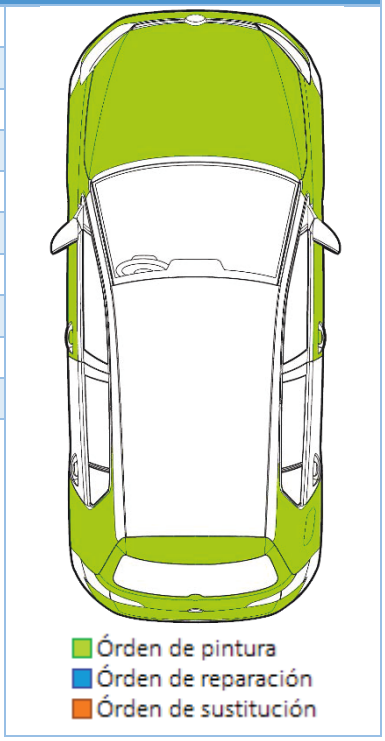


Tabla 14: zona de impacto en siniestros de severidad alta (grupo 2)

Este tipo de siniestro es muy similar, a primera vista, al primer grupo que se ha descrito, ya que las piezas afectadas son similares a las piezas afectadas en este caso, pero en este caso difiere dado que las puertas traseras no son afectadas frecuentemente. Por ello, se puede afirmar que en este clúster se encontrarán todos los siniestros leves con alto número de piezas afectadas en los cuales el vehículo es de tres puertas. Por lo general, serán golpes muy similares a los explicados en el grupo 1.

Grupo 3: golpe lateral derecho

Variables significativas		Porcentaje
SUST_66202	Sustitución de molduras derecha	89,25%
PT_53102	Pintura de aleta trasera derecha	81,69%
PT_57102	Pintura de puerta delantera derecha	80,47%
REP_53102	Reparación de aleta trasera derecha	72,92%
REP_57102	Reparación puerta delantera derecha	72,12%
PT_50102	Pintura de aleta delantera derecha	62,42%
PT_63303	Pintura de paragolpes trasero	56,79%
PT_63203	Pintura de paragolpes delantero	46,25%
PT_58102	Pintura de puerta trasera derecha	43,98%
REP_58102	Reparación de puerta trasera derecha	43,72%

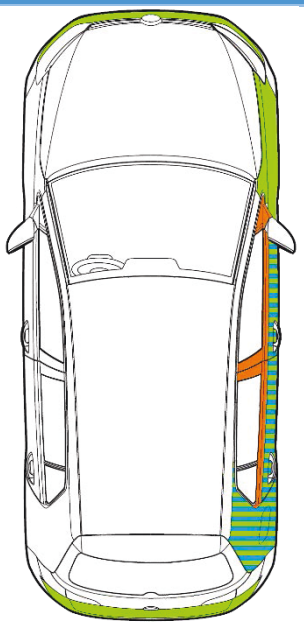


Tabla 15: zona de impacto en siniestros de severidad alta (grupo 3)

En este caso se trata del tipo de golpe más frecuente en siniestros de severidad alta, dado que el 26,04% de siniestros de este tipo se corresponden con esta zona de impacto. Esto podría estar justificado porque el conductor del vehículo está situado en el otro lateral del vehículo y tiene menor control sobre este lateral.

En este grupo se puede observar que existe un daño elevado en las piezas del lateral derecho del vehículo, ocasionando no sólo costes de pintura sino también de reparación de las mimas. Las piezas más afectadas por este tipo de siniestros son las puertas derechas así como la aleta trasera derecha. Otras piezas como la aleta delantera derecha y ambos paragolpes suelen requerir pintura.

Este tipo de siniestro puede darse, por ejemplo, al recibir el vehículo un golpe en un lateral a alta velocidad (en autopista o carreteras convencionales), pero también puede estar ocasionado por una rozadura grave con una columna o una pared, por ejemplo, en un garaje.

Grupo 4: golpe lateral izquierdo

Variables significativas		Porcentaje
SUST_66201	Sustitución de molduras izquierda	85,47%
PT_57101	Pintura de puerta delantera izquierda	83,80%
PT_53101	Pintura de aleta trasera izquierda	77,45%
REP_57101	Reparación puerta delantera izquierda	66,95%
PT_50101	Pintura de aleta delantera izquierda	65,30%
REP_53101	Reparación de aleta trasera izquierda	60,89%
PT_63303	Pintura de paragolpes trasero	56,12%
PT_63203	Pintura de paragolpes delantero	46,27%
PT_58101	Pintura de puerta trasera izquierda	42,15%
REP_50101	Reparación de aleta delantera izquierda	39,51%

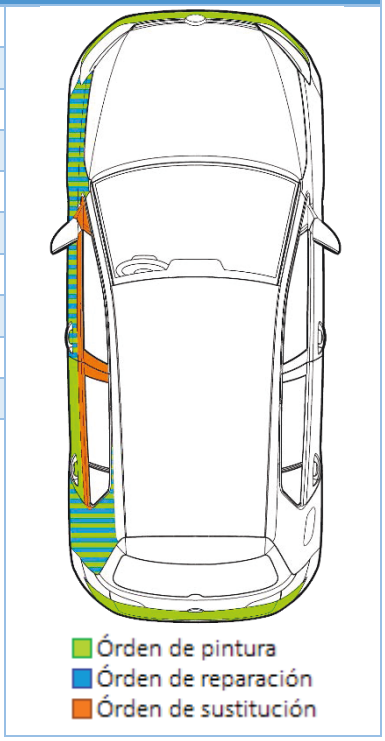


Tabla 16: zona de impacto en siniestros de severidad alta (grupo 4)

Este grupo es muy similar al grupo anterior, pero en este caso los daños son en el lateral opuesto del vehículo. De esta forma, se puede observar que existe un daño elevado en las piezas del lateral derecho del vehículo, ocasionando no sólo costes de pintura sino también de reparación de las mimas. Sin embargo, en el caso de este tipo de golpes existe una ligera diferencia, donde es más frecuente la reparación de la aleta delantera izquierda, y la reparación de la puerta trasera izquierda es menos frecuente.

Al igual que en el caso anterior, este tipo de siniestro puede darse, por ejemplo, al recibir el vehículo un golpe en un lateral a alta velocidad ya sea con otros vehículos, paredes u otros objetos que hubiese en la vía. Otra posible causa de este tipo de accidente por rozaduras profundas con columnas o paredes, las cuales no sean fácilmente reparables.

Grupo 5: golpe general grave (vehículos de 5 puertas)

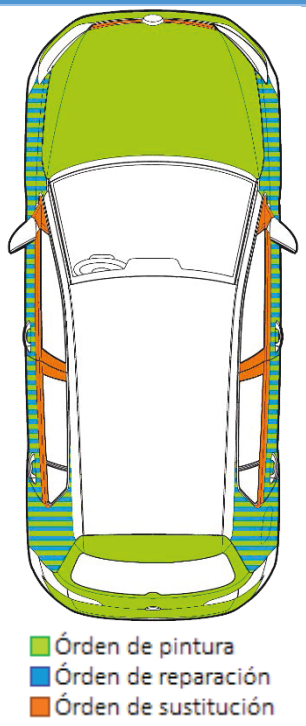
Variables significativas		Porcentaje	
PT_58102	Pintura de puerta trasera derecha	98,42%	
PT_58101	Pintura de puerta trasera izquierda	96,98%	
PT_53102	Pintura de aleta trasera derecha	96,27%	
PT_53101	Pintura de aleta trasera izquierda	95,65%	
REP_58102	Reparación de puerta trasera derecha	93,82%	
PT_57102	Pintura de puerta delantera derecha	90,89%	
REP_58101	Reparación de puerta trasera izquierda	90,16%	
REP_57102	Reparación puerta delantera derecha	88,97%	
PT_63303	Pintura de paragolpes trasero	88,90%	
PT_50101	Pintura de aleta delantera izquierda	87,76%	
REP_57101	Reparación puerta delantera izquierda	87,71%	
PT_55103	Pintura de capó	86,77%	
PT_55303	Pintura de portón trasero	85,22%	
PT_57101	Pintura de puerta delantera izquierda	84,90%	
PT_63203	Pintura de paragolpes delantero	84,81%	
SUST_66202	Sustitución de molduras derecho	83,73%	
PT_50102	Pintura de aleta delantera derecha	83,48%	
SUST_66201	Sustitución de molduras izquierda	82,08%	
REP_53102	Reparación de aleta trasera derecha	80,57%	
SUST_66303	Sustitución de rejilla del radiador	78,74%	
REP_53101	Reparación de aleta trasera izquierda	74,51%	
REP_50101	Reparación de aleta delantera izquierda	73,94%	
REP_50102	Reparación de aleta delantera derecha	72,84%	

Tabla 17: zona de impacto en siniestros de severidad alta (grupo 5)

Este tipo de siniestro incorpora con una frecuencia muy alta órdenes de reparación y pintura en las piezas laterales del vehículo: aletas y puertas, así como órdenes de pintura en las piezas frontales y traseras del vehículo (paragolpes, capó y portón trasero). También incluyen en elevadas ocasiones la necesidad de sustitución de piezas pequeñas (molduras y rejilla del radiador). Se trata, sin duda de siniestros de una gravedad elevada.

En este caso un siniestro de este estilo puede deberse a la pérdida del control del vehículo a gran velocidad, donde el vehículo resulta golpeado contra quitamiedos, paredes u otros obstáculos con los que se haya podido encontrar un vehículo, por ejemplo, con una salida de la vía. Otra posible causa de este tipo de siniestro es un accidente múltiple, donde el vehículo reciba golpes por ambos laterales.

Grupo 6: golpe general grave (vehículos de 3 puertas)

	Variables significativas	Porcentaje
PT_53102	Pintura de aleta trasera derecha	94.72%
PT_53101	Pintura de aleta trasera izquierda	94.62%
REP_53102	Reparación de aleta trasera derecha	91.28%
REP_57102	Reparación puerta delantera derecha	91.16%
REP_57101	Reparación puerta delantera izquierda	90.73%
REP_53101	Reparación de aleta trasera izquierda	89.64%
PT_57102	Pintura de puerta delantera derecha	88.27%
PT_63303	Pintura de paragolpes trasero	86.53%
PT_50101	Pintura de aleta delantera izquierda	84.66%
PT_63203	Pintura de paragolpes delantero	83.71%
SUST_66202	Sustitución de molduras derecha	82.96%
PT_55103	Pintura de capó	81.84%
PT_57101	Pintura de puerta delantera izquierda	81.72%
SUST_66201	Sustitución de molduras izquierda	80.77%
PT_50102	Pintura de aleta delantera derecha	80.50%
PT_55303	Pintura de portón trasero	79.95%
SUST_66303	Sustitución de rejilla del radiador	70.21%

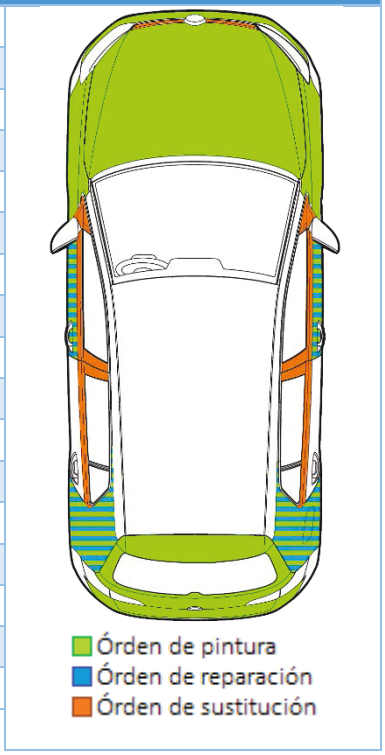


Tabla 18: zona de impacto en siniestros de severidad alta (grupo 6)

En este tipo de siniestros podemos observar como las órdenes más habituales son de pintura en piezas delanteras y traseras del vehículo, incluyendo paragolpes, aletas, capó, portón trasero y las puertas delanteras. De hecho, en las aletas traseras y puertas delanteras se observan también con una frecuencia alta órdenes de reparación. Adicionalmente, se suelen ver afectadas pequeñas piezas que no pueden ser reparadas ni pintadas, sino que son sustituidas, como son las molduras y la rejilla del radiador.

Este tipo de siniestro puede estar ocasionado, por ejemplo, por accidentes en cadena, donde el vehículo afectado colisione tanto por la zona frontal como la zona trasera de manera fuerte. En este tipo de impactos se suelen ver afectadas piezas que rodean la zona de impacto, dado que suelen ser impactos muy fuertes. Este tipo de impacto también puede estar ocasionado por la pérdida de control del vehículo ante un roce con otro vehículo, por las condiciones de la calzada, condiciones meteorológicas o exceso de velocidad.

Grupo 7: golpe trasero

Variables significativas		Porcentaje
SUST_66303	Sustitución de rejilla del radiador	56,61%
PT_63303	Pintura de paragolpes trasero	46,59%
PT_55303	Pintura de portón trasero	34,17%
SUST_66202	Sustitución de molduras derecha	29,47%
SUST_63303	Sustitución de paragolpes trasero	29,35%
SUST_66953	Sustitución de anagrama del fabricante	29,35%
SUST_66201	Sustitución de molduras izquierda	28,43%
PT_53102	Pintura de aleta trasera derecha	28,16%
REP_55303	Reparación de portón trasero	28,16%
REP_63303	Reparación de paragolpes trasero	25,97%

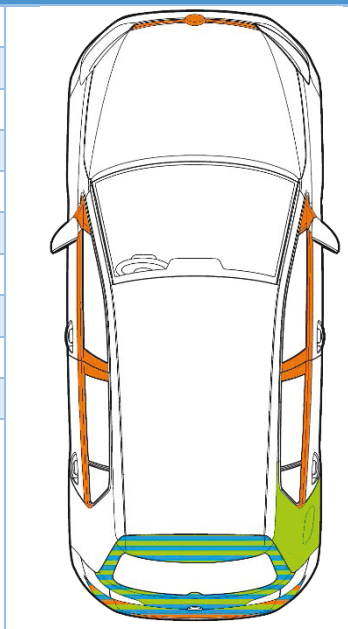


Tabla 19: zona de impacto en siniestros de severidad alta (grupo 7)

En este tipo de golpes se observa una frecuencia de aparición de órdenes de pintura, reparación y sustitución mucho menor que en los demás grupos, por lo que este es el único grupo, de los 7 que hemos visto, en el que las órdenes que aparecen en la tabla anterior tienen menor probabilidad (26 – 57%) de aparecer en un siniestro de este tipo. En este tipo de siniestros se puede observar que los golpes en la zona trasera del vehículo aparecen con bastante frecuencia con órdenes de reparación y de pintura, e incluso con bastante frecuencia con órdenes de reparación. También podemos observar la sustitución de otras piezas de menor valor, como son el anagrama del fabricante, la rejilla del radiador o las molduras de ambos laterales del vehículo, con cierta frecuencia. En este caso se considera que este grupo engloba los golpes traseros, pero que las consecuencias de este tipo de golpes son más difíciles de predecir que en los otros casos.

Estos golpes estarán ocasionados, por lo general, por alcances a alta velocidad, por otro tipo de impactos mientras el vehículo se mueve hacia atrás o mientras el vehículo está estacionado y otros vehículos circulan a su alrededor a una velocidad media o alta.

4.1.2.3. Grupo 3: severidad media con gasto elevado en piezas

Se han obtenido ciertas estadísticas que representan los datos contenidos en este clúster, como se puede observar en la siguiente tabla.

	Tot_mo *Coste de mano de obra*	Tot_pint *Coste de pintura de piezas*	Tot_sust *Coste de piezas sustituidas*
Mínimo	0,00 €	0,00 €	345,80 €
Máximo	1.170,70 €	1.090,90 €	1.553,50 €
Media	344,32 €	301,99 €	826,13 €
Cuartil 1	187,07 €	209,96 €	601,55 €
Mediana	282,48 €	299,03 €	771,22 €
Cuartil 3	448,34 €	384,07 €	1.016,00 €
Desv. Típica	215,09 €	162,85 €	269,90 €

Tabla 20: distribución de costes en siniestros de severidad media con gasto en piezas

Por otra parte, en el clúster número 3 podemos ver que el coste de sustitución de piezas es intermedio, en el punto exacto entre los dos clúster anteriormente mencionados. Por otra parte, se puede observar que el coste de pintura es por lo general bajo, lo cual se ve también en la media, mediana y cuartiles de la muestra, los cuales son también bajos y tienen una desviación típica baja. En el caso del coste de mano de obra, se puede ver como por lo general es medio-bajo pero con una desviación un poco más alta. Por tanto, este clúster engloba todos aquellos siniestros de gravedad media en los cuales las piezas por lo general no han podido ser reparadas y han tenido que ser sustituidas. Este tipo de siniestros podría englobar, por ejemplo, un golpe contra una pared que haya ocasionado daños graves a las piezas. Por lo anteriormente mencionado, este clúster será denominado como *Severidad media con gasto elevado en piezas*.

A continuación se ha realizado un análisis de correlaciones entre las variables que se están estudiando en este clúster, el cual nos indicará cómo están relacionadas entre sí estas variables.

	Tot_mo *Coste de mano de obra*	Tot_pint *Coste de pintura de piezas*	Tot_sust *Coste de piezas sustituidas*
Tot_mo *Coste de mano de obra*	1	0,5301	0,1348
Tot_pint *Coste de pintura de piezas*	0,5301	1	0,0925
Tot_sust *Coste de piezas sustituidas*	0,1348	0,0925	1

Tabla 21: correlación entre los costes en siniestros de severidad media con gasto en piezas

En esta tabla, al igual que en los dos grupos anteriores, se puede observar cómo se ha reducido en gran medida la correlación entre todas estas variables con respecto a la base de datos completa. Se puede ver cómo al igual que en los casos anteriores sigue existiendo una correlación positiva y moderada entre el coste de pintura y el coste de mano de obra, lo que indica que estas variables están bastante influenciadas: el incremento de costes de mano de obra suele implicar un incremento en costes de pintura y viceversa. Adicionalmente, se puede observar una correlación muy baja entre los costes de sustitución de piezas y los demás costes, lo cual indica que estas variables no son muy dependientes entre sí. En concreto se puede observar que los

costes de mano de obra y los costes de sustitución de piezas están correlacionados de forma muy baja y positiva, de manera que la relación es lineal (si unos costes suben, los otros también pero en una medida muy reducida). Esto mismo ocurre entre los costes de sustitución de piezas y los costes de pintura.

Tras ejecutar el proceso con este tipo de siniestros, se han podido encontrar dos grupos diferenciados de zonas de impacto. Como se puede observar en la siguiente gráfica, existe una proporción similar de siniestros en cada uno de los clústeres.

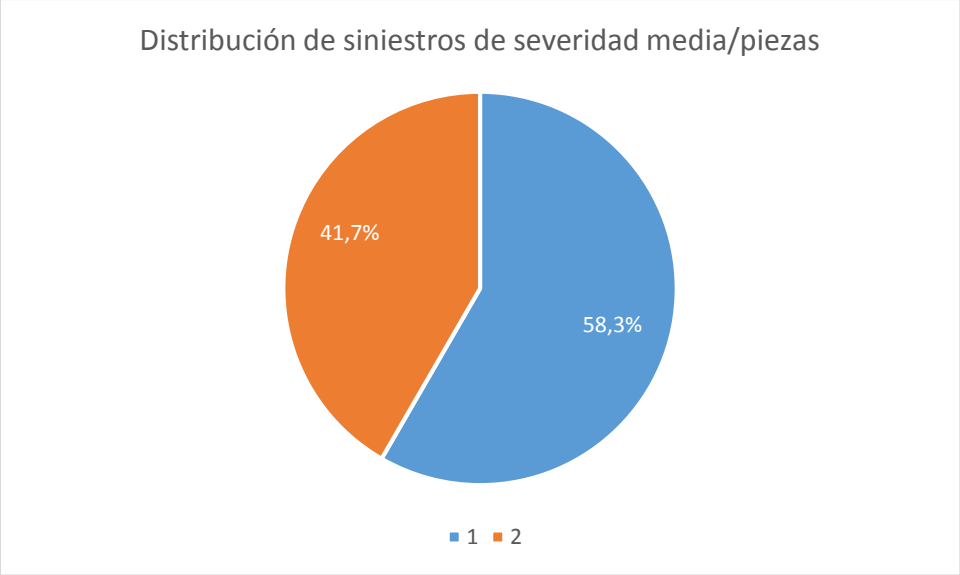


Ilustración 26: distribución de los siniestros de severidad media con gasto elevado en piezas

Dentro de ambos grupos se pueden encontrar órdenes de sustitución y pintura con una frecuencia elevada, especialmente en el primer clúster. Como se verá en la tabla a continuación, existen dos tipos de golpes en este grupo de severidad: golpes frontales y golpes traseros.

Clúster	Siniestros	Variables significativas		Porcentaje
1: Golpe frontal	58,35%	PT_63203	Pintura de paragolpes delantero	95,09%
		SUST_63203	Sustitución de paragolpes delantero	91,66%
		SUST_50203	Sustitución componentes de la coraza	67,46%
		PT_55103	Pintura del capó	62,34%
		SUST_94101	Sustitución de faro delantero derecho	57,05%
		SUST_94102	Sustitución de faro delantero izquierdo	53,88%
		SUST_66953	Sustitución de anagrama del fabricante	47,53%
		SUST_63202	Sustitución paragolpes delantero derecho	45,87%
		SUST_63201	Sustitución paragolpes delantero izquierdo	44,71%
		SUST_66103	Sustitución de luna custodia	43,63%

Tabla 22: zonas de impacto en siniestros de severidad media con gasto elevado en piezas (I)

Clúster	Siniestros	Variables significativas		Porcentaje
2: Golpe trasero	41,65%	PT_63303	Pintura de paragolpes trasero	42,22%
		SUST_63303	Sustitución de paragolpes trasero	29,85%
		SUST_66201	Sustitución de molduras izquierda	26,87%
		PT_53101	Pintura de aleta trasera izquierda	25,56%
		SUST_66202	Sustitución de molduras derecha	25,56%
		PT_57101	Pintura de puerta delantera izquierda	24,68%
		PT_53102	Pintura de aleta trasera derecha	24,46%
		PT_57102	Pintura de puerta delantera derecha	23,35%
		PT_63203	Pintura de paragolpes delantero	21,79%
		PT_55303	Pintura de portón trasero	19,73%

Tabla 23: zonas de impacto en siniestros de severidad media con gasto elevado en piezas (II)

En este caso se observa en primer lugar que los golpes delanteros presentan una homogeneidad mucho más alta en los siniestros que contiene en su interior, ya que hay varias órdenes de pintura y sustitución de piezas que tienen una frecuencia muy elevada. En este caso, es habitual la sustitución y pintura del paragolpes delantero, lo cual como se comentaba con anterioridad es habitual dado que el paragolpes está situado sobresaliendo del vehículo para absorber los golpes frontales. También se presentan otras órdenes habituales como sustitución de piezas de la zona delantera (coraza, faros, elementos del paragolpes, etcétera). El hecho de que estos porcentajes sean tan altos significa que, probablemente, en los siniestros acaecidos se han visto implicadas varias piezas al mismo tiempo.

Por otra parte, los golpes traseros muestran órdenes de pintura o sustitución que se repiten con una menor frecuencia, aunque esta frecuencia sigue siendo alta. Lo más frecuente, en este caso, es la pintura y/o sustitución del paragolpes trasero, así como la pintura de las aletas y sustitución de molduras. En este caso, también se ven afectadas habitualmente las puertas del vehículo, lo cual indica que este tipo de golpes suele abarcar un área más grande que los golpes frontales.

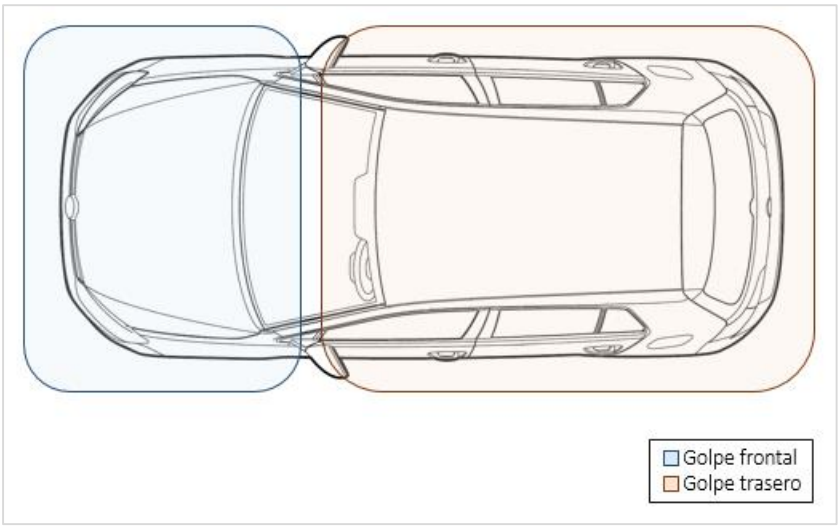


Ilustración 27: zonas de impacto en siniestros de severidad media con gasto alto en piezas

En la ilustración anterior se pueden observar, representadas gráficamente, las zonas de impacto definidas previamente.

Al igual que en los siniestros de severidad media con gasto elevado en pintura, se puede observar como las piezas afectadas son de más elevado coste de reparación, pero siguen siendo piezas bastante superficiales, por lo que se confirma que se trata de siniestros de severidad media.

4.1.2.4. Grupo 4: severidad media con gasto elevado en pintura

Se han obtenido ciertas estadísticas que representan los datos contenidos en este clúster, como se puede observar en la siguiente tabla.

	Tot_mo *Coste de mano de obra*	Tot_pint *Coste de pintura de piezas*	Tot_sust *Coste de piezas sustituidas*
Mínimo	0,00 €	123,30 €	0,00 €
Máximo	1.076,50 €	1.091,80 €	777,65 €
Media	349,97 €	673,48 €	134,98 €
Cuartil 1	225,21 €	487,26 €	23,39 €
Mediana	319,78 €	671,11 €	61,65 €
Cuartil 3	444,00 €	855,44 €	227,28 €
Desv. Típica	173,94 €	208,49 €	145,61 €

Tabla 24: distribución de costes en siniestros de severidad media con gasto en pintura

Por último, el grupo restante, muestra un coste de mano de obra muy similar al del grupo anterior, también con una dispersión media. En el caso de los costes de pintura, en este caso son por lo general altos, superando los costes medios así como los cuartiles y la mediana de los anteriores grupos, y con una dispersión media-alta. En este caso, el coste de sustitución de piezas es, por lo general, bajo y con una desviación típica media. Todo esto nos indica que este tipo de siniestro no suele implicar sustitución de gran cantidad de piezas, sino que las piezas dañadas suelen ser reparables. Un ejemplo de este siniestro podría ser un raspón de un lateral completo de un vehículo contra un muro. En este caso, se denominará al grupo *Severidad media con gasto elevado en pintura*.

A continuación se ha realizado un análisis de correlaciones entre las variables que se están estudiando en este clúster, el cual nos indicará cómo están relacionadas entre sí estas variables.

	Tot_mo *Coste de mano de obra*	Tot_pint *Coste de pintura de piezas*	Tot_sust *Coste de piezas sustituidas*
Tot_mo *Coste de mano de obra*	1	0,0168	0,3400
Tot_pint *Coste de pintura de piezas*	0,0168	1	0,0079
Tot_sust *Coste de piezas sustituidas*	0,3400	0,0079	1

Tabla 25: correlación entre los costes en siniestros de severidad media con gasto en pintura

Esta tabla difiere en gran medida con la de los grupos anteriores, ya que muestra las correlaciones más bajas de todos los casos anteriores con respecto a la base de datos completa. Adicionalmente, se puede observar una correlación muy baja entre los costes de sustitución de piezas y los demás costes, lo cual indica que estas variables no son muy dependientes entre sí. En concreto se puede observar que los costes de pintura tienen una correlación muy baja, casi cero, lo que indica que variaciones en esta variable casi no influyen sobre las demás y viceversa. Sin embargo, entre los costes de sustitución de piezas y los costes de mano de obra existe una correlación positiva y moderada, lo que indica que estas variables están influenciadas: el incremento de costes de sustitución de piezas suelen implicar un incremento en costes de mano de obra y viceversa, lo cual tiene sentido ya que generalmente se cobrarán al cliente gastos de mano de obra al reemplazar piezas a su vehículo.

En este caso no ha sido posible realizar categorizaciones en grupos con consistencia. Por una parte, la ejecución del proceso *Canopy* recomienda utilizar un único clúster y, por otra parte, las pruebas que se han realizado con 2, 3 y 4 grupos dan unos resultados que no aportan información válida. Las pruebas que se han realizado con varios grupos muestran un grupo con características similares entre todas ellas, y los siniestros del grupo o los grupos adicionales tienen muy poca relación entre sí, mostrando una repetición máxima de una orden de reparación, pintura o sustitución inferiores al 7% en todos los casos.

Por todo ello, se puede afirmar que en este tipo de siniestros no existe un patrón común, ni tampoco existe ningún grupo claramente formado. Se ha construido la siguiente tabla basándose en el análisis realizado sobre todos los datos sin realizar separación por grupos.

Variables significativas		Porcentaje
PT_63203	Pintura paragolpes delantero	25,13%
SUST_63203	Sustitución paragolpes delantero	23,60%
SUST_50203	Sustitución componentes de la coraza	20,47%
PT_55103	Pintura capó	18,60%
SUST_94101	Sustitución faro delantero izquierdo	17,23%
SUST_94102	Sustitución faro delantero derecho	17,02%
SUST_63202	Sustitución de paragolpes delantero derecho	16,39%
SUST_63201	Sustitución de paragolpes delantero izquierdo	15,96%
PT_50101	Pintura de aleta delantera izquierda	15,29%
PT_50102	Pintura de aleta delantera derecha	15,10%

Tabla 26: zonas de impacto en siniestros de severidad media con gasto elevado en pintura

Como se puede observar en la tabla anterior, las órdenes de pintura o sustitución más habituales son en la parte delantera del vehículo, abarcando el paragolpes, el capó, los faros y las aletas delanteras. En comparación con los siniestros de severidad baja, se puede observar como las piezas afectadas no son superficiales, a diferencia de los anteriores, y la sustitución de piezas es

más habitual que la pintura, por lo que se ve como los siniestros son de una severidad mayor, aunque las piezas que son arregladas siguen sin ser de muy costosa reparación.

4.1.3. Conclusiones sobre los grupos de severidad obtenidos

Una vez analizados a grandes rasgos los grupos que hemos obtenido del proceso, resulta interesante representar gráficamente la distribución de cada una de las variables en los distintos clúster, para profundizar en el análisis de los datos obtenidos. En la siguiente ilustración se pueden observar las distribuciones de estas tres variables en los distintos clúster.

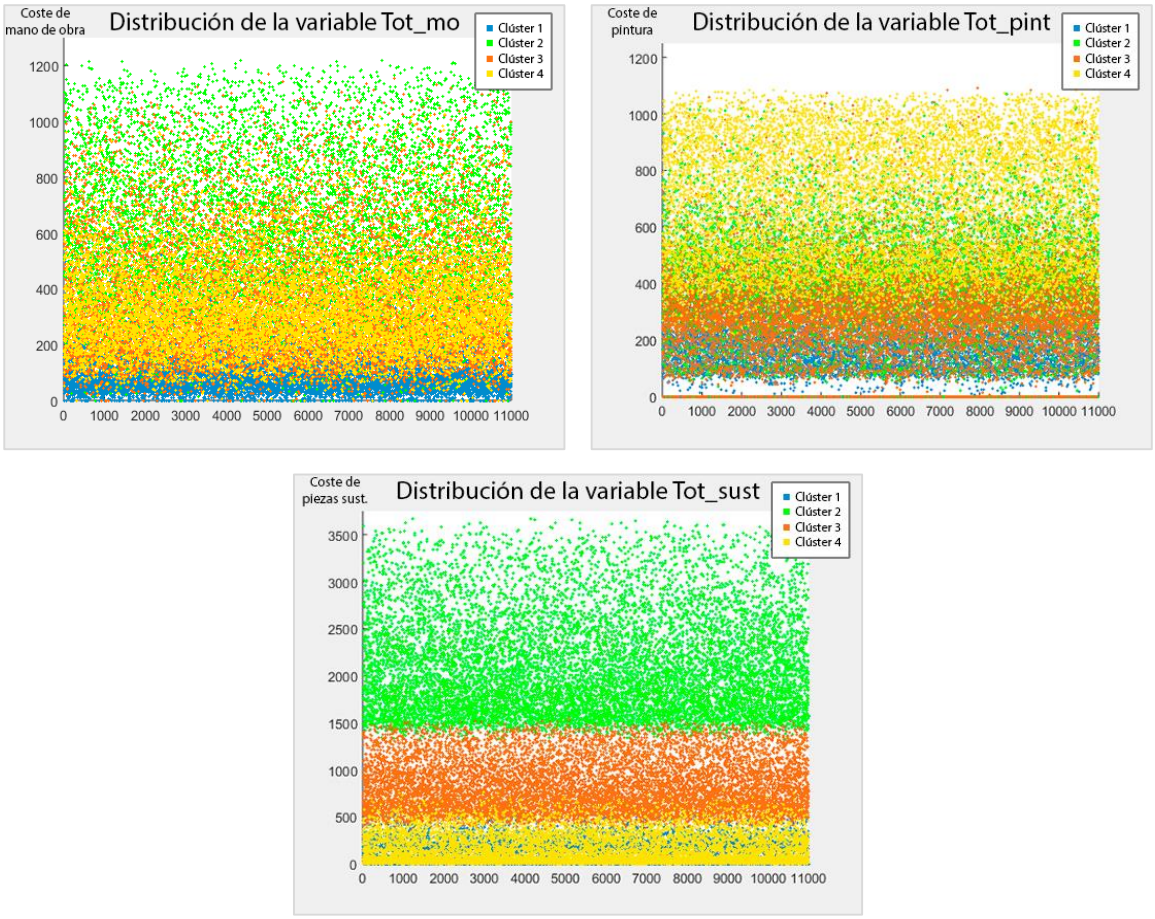


Ilustración 28: distribución de las variables Tot_mo, Tot_pint y Tot_sust por separado

Como se puede observar en las gráficas anteriores, se confirman los datos obtenidos con anterioridad. En primer lugar, se puede observar como el primer clúster se encuentra en los rangos más bajos de precios en los tres casos. De la misma manera, el segundo clúster abarca franjas de una altura superior, es decir, los valores están más dispersos, y siempre se sitúa por los valores más altos. Por otra parte, se reconfirma que los clúster de severidad media (2 y 3) tienen unos costes de mano de obra muy similares mientras que difieren en los costes de pintura y de piezas sustituidas.

Podemos observar que en el primer clúster (siniestros de severidad baja), los costes son siempre bajos, y tienen una amplitud muy inferior a la del resto de grupos. Esto reconfirma que se trata

de los siniestros con menor coste económico para la aseguradora. El resto de grupos, sin embargo, tienen una amplitud mucho mayor, lo cual nos dice que el coste de un nuevo siniestro de este tipo no está tan acotado.

En la primera gráfica de la ilustración anterior se puede observar que el coste de mano de obra es muy disperso, y los grupos no quedan claramente definidos. Para una aseguradora, esto significa que ante un nuevo siniestro de cualquier severidad, el coste de mano de obra no es fácilmente predecible.

Por último, se puede observar que la variable *Tot_sust* tiene un papel importante en la formación de los grupos, dado que no hay apenas superposición de valores en las franjas en las que se divide (exceptuando los clúster 1 y 4, que toman valores muy similares).

También resulta interesante representar la distribución de estas variables enfrentándolas dos a dos, para ver cómo están relacionadas las variables entre sí en cada uno de los clúster.

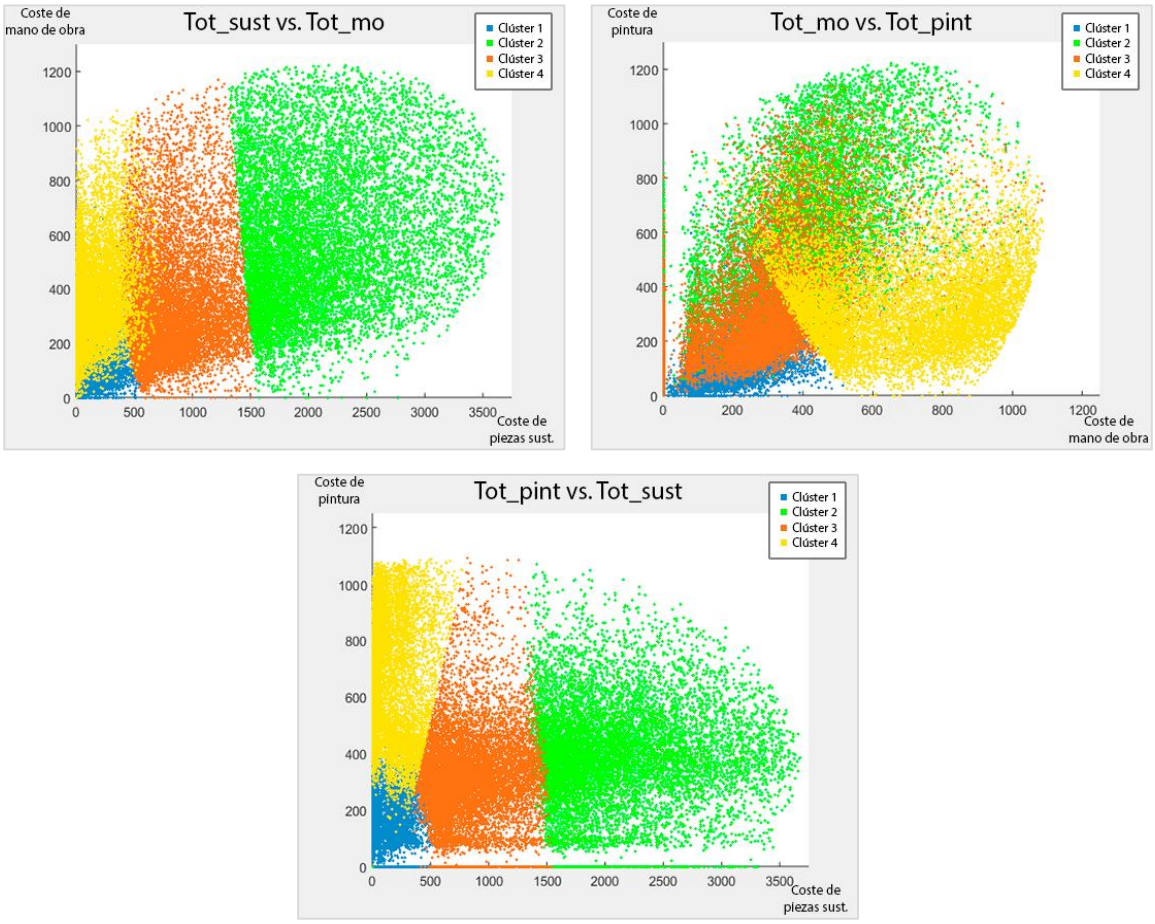


Ilustración 29: distribución comparada de las variables Tot_mo, Tot_pint y Tot_sust

Se puede observar claramente como el coste de pintura y el de mano de obra tienen zonas muy definidas sin apenas superposición de observaciones. Lo mismo ocurre con la comparación entre el coste de pintura y el coste de piezas sustituidas. Sin embargo, comparando el coste de pintura

y el de mano de obra podemos observar como no existe una separación tan clara, y los límites de cada clúster no están muy bien definidos, en especial con respecto al clúster número dos, el cual es mucho más disperso.

Se ha realizado un esquema que recoge todos los grupos de severidad y subgrupos de zonas de impacto que se han creado en este apartado, el cual se puede observar en la ilustración a continuación.

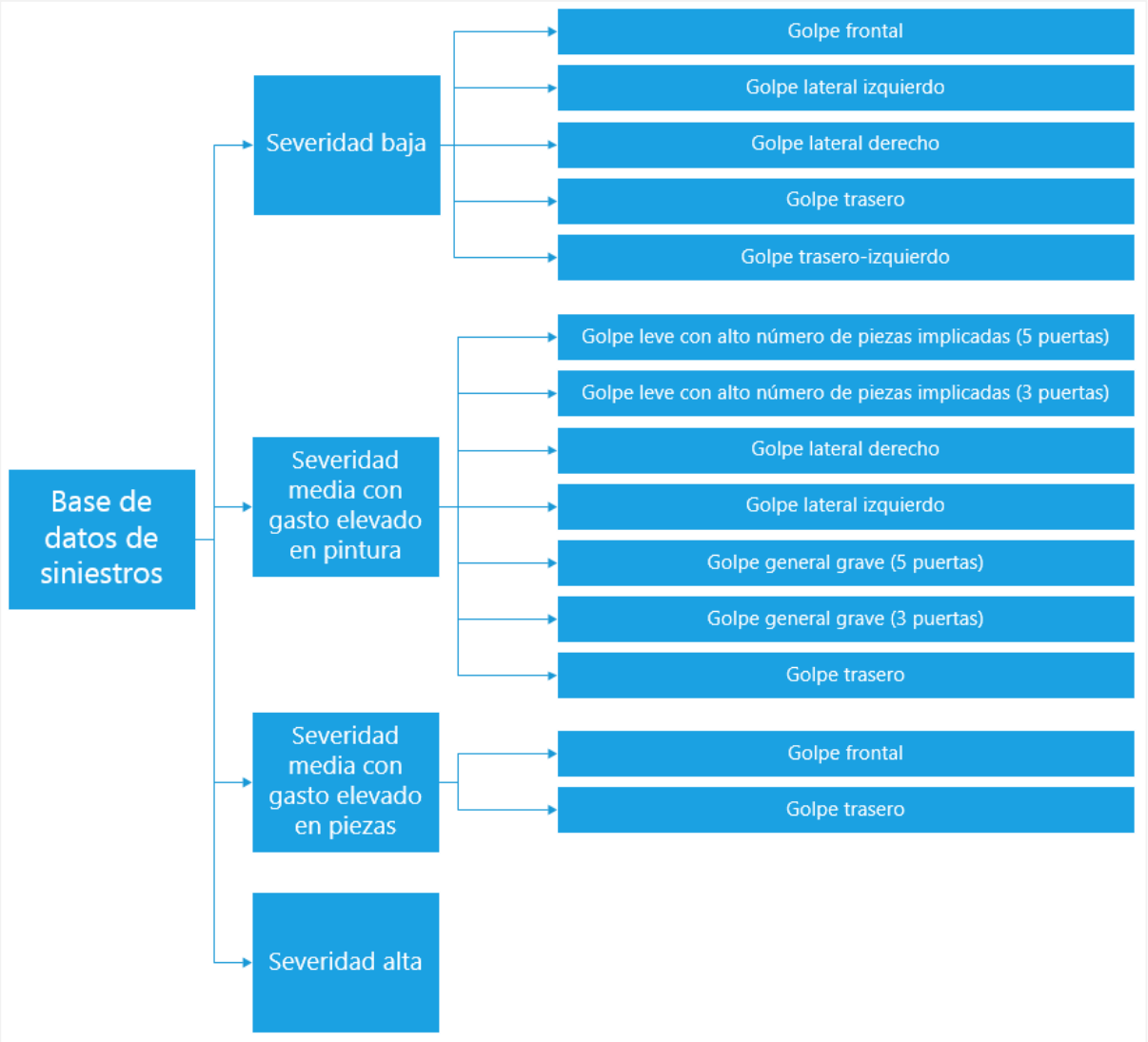


Ilustración 30: clasificación de los siniestros por severidad

Por último, se han resumido los datos comentados anteriormente en la siguiente tabla, que muestra los clúster que se han obtenido así como el valor que toma cada una de las variables en ellos y la dispersión de los mismos.

Clúster		Coste de mano de obra	Coste de pintura	Coste de piezas sustituidas
Severidad baja (1)	Valor	Bajo	Bajo	Bajo
	Dispersión	Baja	Baja	Baja
	Rango	[0 – 827,33]	[0 – 526,77]	[0 – 582,63]
	Media	99,69€	157,58€	102,71€
Severidad media con gasto elevado en pintura (4)	Valor	Medio	Alto	Bajo
	Dispersión	Media	Media-alta	Media
	Rango	[0 – 1.076,5]	[123,3– 1.091,8]	[0 – 777,65]
	Media	349,97€	673,48€	134,98€
Severidad media con gasto elevado en piezas (3)	Valor	Medio-bajo	Bajo	Medio
	Dispersión	Media	Baja	Media
	Rango	[0 – 1.170,7]	[0 – 1.090,9]	[348,8 – 1.553,5]
	Media	344,32€	301,99€	826,13€
Severidad alta (2)	Valor	Alto	Alto	Alto
	Dispersión	Alta	Media	Alta
	Rango	[0 – 24.800]	[0 – 1.069,5]	[0 – 3.672,3]
	Media	583,2€	387,62€	2.083,4€

Tabla 27: descripción de los clústeres de severidad obtenidos

4.1.4. Creación de árboles de decisión

Una vez generados los grupos de severidad, es el momento de crear árboles de clasificación, que sean capaces de categorizar nuevos siniestros con respecto a este vehículo, en cualquiera de los cuatro grupos de severidad de primer nivel. Para ello, como se ha comentado con anterioridad, se utilizará el algoritmo *Random Forests* incluido en Mahout.

Tras realizar el preprocesado necesario, se dispone de la base de datos de siniestros, con una columna extra que incluye una nueva variable, la cual toma valores de 1 a 4 dependiendo del grupo al que pertenezca la observación. Se ha ejecutado el algoritmo de creación de árboles de decisión proporcionándole las siguientes cuatro columnas:

- Tot_mo: total de costes de mano de obra del siniestro (numérica).
- Tot_pint: total de costes de pintura del siniestro (numérica).
- Tot_sust: total de costes de sustitución de piezas (numérica).
- Clúster: número identificador del clúster de severidad de primer nivel al que pertenece el siniestro (etiqueta).

La ejecución de este proceso se realiza en un tiempo muy reducido, apenas 3 minutos y medio. El resultado del mismo se puede observar en la ilustración a continuación.

```
15/09/18 23:49:22 INFO common.HadoopUtil: Deleting hdfs://giaa-edge1:54310/user/hadoop/ns1-forest
15/09/18 23:49:22 INFO mapreduce.BuildForest: Build Time: 0h 3m 27s 9
15/09/18 23:49:22 INFO mapreduce.BuildForest: Forest num Nodes: 41316
15/09/18 23:49:22 INFO mapreduce.BuildForest: Forest mean num Nodes: 413
15/09/18 23:49:22 INFO mapreduce.BuildForest: Forest mean max Depth: 14
15/09/18 23:49:22 INFO mapreduce.BuildForest: Storing the forest in: ns1-forest/forest.seq
15/09/18 23:49:22 INFO driver.MahoutDriver: Program took 208126 ms (Minutes: 3.4687666666666668)
hadoop@giaa-edge1:~/TFG/Análisis/3. Arboles/1. Severidad$
```

Ilustración 31: información sobre los árboles de decisión de severidad

Se puede observar que se ha generado un bosque con 41.316 nodos, donde la longitud máxima de un árbol es de 14 nodos, y la media de nodos por árbol es de 413. Esto nos dice que existen un total de 100 árboles, los cuales tienen una amplitud (número de hijos por nodo) muy elevada y una profundidad moderada. Esto demuestra la complejidad del bosque generado.

Dada la complejidad de este bosque, es de gran importancia poner a prueba dichos árboles para asegurar que su fiabilidad es máxima. Una vez realizadas las pruebas, de un total de 92.706 observaciones de prueba, 92.438 han sido clasificadas correctamente, lo cual supone un 99.71% de precisión, un valor muy elevado. A continuación se puede observar la matriz de confusión, la cual nos permite desglosar estas pruebas.

	Clasificado en cl. 1	Clasificado en cl. 2	Clasificado en cl. 3	Clasificado en cl. 4
Pertenece a cl. 1	**67.005**	0	22	14
Pertenece a cl. 2	1	**3.278**	34	0
Pertenece a cl. 3	65	5	**10.716**	18
Pertenece a cl. 4	89	0	20	**11.439**
Tasa de acierto	*99,77%*	*99,85%*	*99,30%*	*99,72%*

Tabla 28: matriz de confusión de los árboles de grupos de severidad

En esta tabla se puede observar, para cada clúster, el número de observaciones que fueron clasificadas en él (columnas) y el número de observaciones que pertenecen a cada clúster (filas).

De esta forma, se puede observar el número de siniestros clasificados correctamente en cada clúster, en la línea diagonal. Es importante observar que, por ejemplo, casi no se han asignado erróneamente observaciones en el clúster número 2 (severidad alta). Además, se puede observar también que el clúster 1 es en el que más instancias se han asignado de forma incorrecta.

A continuación se pueden observar las variables estadísticas que se pueden extraer a partir del resultado de las pruebas realizadas.

Kappa	-0,8008
Precisión	99,71%
Fiabilidad	79,42%
σ (fiabilidad)	0,4440

Tabla 29: variables estadísticas de los árboles de grupos de severidad

Como se puede observar, la precisión de este bosque es muy elevada, dando resultados de clasificación de siniestros buenísimos. El valor de *Kappa*, al ser negativo, indica que en este caso la precisión obtenida es superior a la fiabilidad esperada. La fiabilidad de este bosque es muy alta (79,42%), además de que su desviación típica es muy reducida, lo que demuestra que se trata de un modelo muy fiable y excepcionalmente preciso.

4.2. Creación de grupos de zonas de impacto

Un punto clave en el estudio de siniestros en el sector automóvil es averiguar qué piezas de los vehículos son afectadas cuando se producen siniestros, es decir, qué zonas del vehículo son dañadas dependiendo del tipo de impacto que se haya producido. Para una compañía aseguradora conocer esta información puede ser de gran ayuda para prever los costes que puede generar un siniestro. Adicionalmente, esto puede ayudar en gran medida a evitar el fraude en seguros de automóviles, pudiendo detectar de forma fácil siniestros donde se hayan incluido más piezas a reparar de las que realmente han sido afectadas por el siniestro.

Además, se puede profundizar en este análisis si dentro de cada grupo resultante de zonas de impacto se agrupan los datos según la severidad, ya que esto sería muy interesante para que la compañía pudiese estimar, partiendo de la zona que ha sido impactada, el coste total de la reparación del siniestro.

4.2.1. Proceso de obtención de las zonas de impacto

De este análisis se pretenden extraer varias agrupaciones de piezas afectadas por un siniestro que determinen las diferentes zonas de impacto que pueden darse en un siniestro. A partir de estas zonas se podrá concluir qué piezas son las más comúnmente afectadas por este tipo de golpe, y así formar una serie de patrones sobre la forma que tienen los siniestros.

En un primer lugar, se ha intentado llevar a cabo este análisis con las 710 variables binarias que indican qué piezas han sido reparadas, qué piezas han sido pintadas y cuáles han sido sustituidas. Este análisis ha sido fallido con todos los números de clústeres probados (de 4 a 20) y con todas las medidas de distancia explicadas con anterioridad. Los clústeres resultantes no sólo tenían observaciones que apenas se relacionaban entre sí, sino que las piezas más afectadas, al ser dibujadas sobre un vehículo, no tenían ningún sentido, ya que afectaban a partes del vehículo que estaban muy alejadas, o todos los clúster mostraban observaciones muy similares. Esto ha hecho que sea necesario replantearse el análisis que se va a realizar y el motivo por el que esto ocurre.

Para una base de datos como la que nos ocupa, y con un número de variables tan grande, es normal que los valores sean complicados de clasificar. Esto es debido a que disponemos de una cantidad muy grande de datos que no han sido previamente clasificados, por lo que se trata de observaciones aleatorias con muchos parámetros, en los cuales son difícilmente identificables

ningún tipo de patrones. En el caso de la obtención de zonas de impacto sobre los grupos de severidad que se ha realizado con anterioridad, el resultado era claro porque los siniestros de cada uno de esos grupos de severidad tenían ciertas características comunes entre sí, pero en este caso disponemos de la totalidad los siniestros, sin tener características comunes.

Por lo anteriormente mencionado, se ha tomado la decisión de simplificar la base de datos de entrada para poder disponer de una cantidad de variables más pequeña pero muy precisa. En este caso, se ha decidido omitir el tipo de trabajo que se ha realizado sobre la pieza, teniendo en cuenta únicamente las piezas que han sido afectadas. Por ello, para cada pieza se han creado nuevas variables en la base de datos de tipo binario, que contiene únicamente los códigos de las piezas, y toman el valor 1 si han sido dañadas en el siniestro, es decir, `XXXX` toma el valor 1 si al menos una de las variables `PT_XXXXX`, `REP_XXXXX` o `SUST_XXXXX` tiene el valor 1, siendo `XXXXX` el código de pieza. En caso contrario estas variables tomarán el valor 0. Adicionalmente, se han omitido de estas nuevas variables aquellas piezas cuyo número de apariciones no era significativa, es decir, aquellas cuyo número de apariciones sea inferior al 0,1% de la muestra, ya que no se consideran significativas para el estudio, y pueden introducir ruido. Las variables tenidas en cuenta se pueden observar en las tablas 70 y 71 (Anexo II).

Una vez creadas las nuevas variables para el estudio, se proporcionarán a Mahout estas 75 variables binarias recientemente creadas para este estudio para realizar la separación en clústeres.

En primer lugar, se ha realizado el preprocesado de datos, pero no ha sido necesario realizar ningún tipo de normalización a los mismos ya que el proceso no realizaría ningún cambio en los datos, al tener todos valores 0 o 1. En el caso de la eliminación de datos atípicos, se han eliminado un total de 176 observaciones atípicas, lo cual supone el 0,053% de los datos, lo cual es una cantidad de datos atípicos muy reducida.

Para continuar, como se ha venido realizando en todos los análisis realizados con anterioridad, se ejecutará el algoritmo *Canopy* sobre estas variables. Se han elegido los siguientes parámetros para este algoritmo:

- *Método de medida de distancia*: de la misma manera que en el apartado 2.6.2.3, se ha decidido emplear la distancia *Euclídea*, ya que se ha observado que es la que proporciona resultados con una mejor relación intra-clúster este análisis.
- *Umbral 1*: se ha establecido como el valor 50. De esta forma, serán considerados para formar parte de un clúster todos aquellos siniestros cuya distancia al centroide del clúster sea menor que este número. Este número permitirá al proceso *Canopy* considerar para

formar parte de un clúster a vectores cuya distancia sea bastante grande sin comprometer el tiempo de respuesta del programa.

- *Umbral 2*: se ha establecido como 10. De esta forma, todas aquellas observaciones cuya distancia al centroide de un grupo sea inferior a este valor serán consideradas como pertenecientes a un clúster y no serán movidas, ya que se considera que una diferencia de 10 piezas afectadas entre los distintos siniestros que se hallen en un grupo algo muy pequeño.

En este caso, al igual que en el análisis de zonas de impacto realizado con anterioridad, dado que tenemos un número muy grande de variables, y el algoritmo *Canopy* es bastante sensible a los umbrales elegidos, con el fin de eliminar los posibles fallos introducidos por la elección de estos valores se utilizará el número de clústeres que resulten como aproximado, y se repetirá el método *k-Means* en diversas ocasiones, variando el número de grupos a obtener para afinar el resultado entre un número de clústeres cercanos al obtenido. De esta forma, si al variar el número de grupos se observa que uno de los grupos da resultados con poca relación entre ellos, o con una cantidad muy pequeña de observaciones en su interior, sabremos que dicho número de grupos no es válido.

Una vez obtenidas las diferentes zonas de impacto, resulta muy interesante poder analizar qué grupos de severidad podemos encontrar dentro de cada uno de ellos, es decir, poder crear distintos grupos de gravedad de los siniestros partiendo de la zona de impacto del mismo. Esto permitirá relacionar el análisis hecho recientemente con el impacto económico de estos siniestros, lo cual es tremendamente útil para, por ejemplo, predecir a partir de una zona de impacto el coste económico que puede acarrear cada siniestro.

Para realizar este análisis se ejecutará el algoritmo *Canopy* sobre estas variables, de la manera que se ha explicado con anterioridad, utilizando las variables `Tot_mo`, `Tot_pint` y `Tot_sust`, las cuales nos indican los costes totales de mano de obra, pintura y sustitución de piezas respectivamente.

Se han elegido los mismos parámetros que en el apartado 4.1.1 para analizar cada uno de los siete grupos de zonas de impacto que se han hallado con anterioridad, ya que se trata de un análisis de idénticas características, a diferencia que los datos de entrada, en este caso, serán mucho más similares entre sí que en el caso anterior. Los parámetros introducidos son los siguientes:

- *Método de medida de distancia*: distancia Euclídea.
- *Umbral 1*: 3.500,00€.
- *Umbral 2*: 100,00€

Una vez obtenido el número de grupos, se ejecutará sobre estos datos el algoritmo *k-Means* utilizando como clústeres de partida los obtenidos por el algoritmo *Canopy*, el mismo método de medida de distancia que anteriormente y el número de grupos obtenidos por el algoritmo *Canopy*.

4.2.2. Análisis de las zonas de impacto obtenidas

Tras ejecutar el algoritmo probando con distinto número de clústeres (de 6 a 9), se ha determinado que el resultado más preciso es el de 7 clústeres, dado que los demás resultados mostraban o bien grupos demasiado similares entre sí (probablemente repetidos), o bien mostraban varios clústeres de un tamaño grande y otros de un tamaño muy reducido para la muestra que se está analizando. Por ejemplo, con 9 clústeres se podía observar un grupo de 155 observaciones, el cual no es significativo para la cantidad de datos que estamos analizando.

En primer lugar, se ha realizado un gráfico donde se puede observar la distribución de estas piezas en diferentes clústeres.

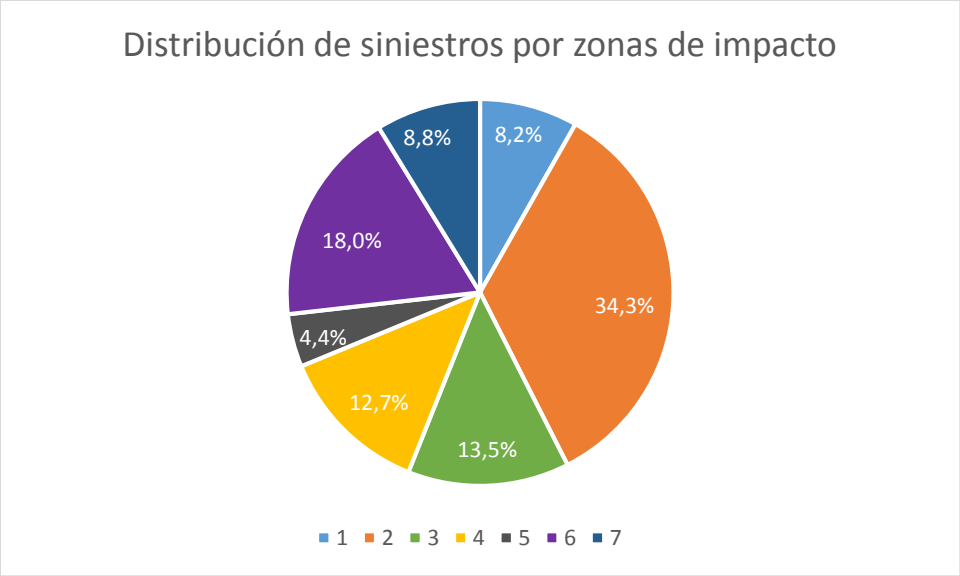


Ilustración 32: distribución de los siniestros por zonas de impacto

A continuación se describirán de forma detallada cada uno de los 7 clústeres obtenidos, explicando las piezas que son afectadas con más frecuencia y dibujándolas sobre un boceto de un vehículo. Se considerará que una pieza es afectada con frecuencia cuando al menos el 40% de las observaciones en su interior hayan tenido que realizar trabajos sobre ella. Este porcentaje ha sido elegido dado que, como se comentó con anterioridad, estas observaciones no han sido previamente clasificadas y, por tanto, serán muy diferentes entre ellas. En todos los casos, se detallarán al menos las 10 piezas afectadas más frecuentemente.

4.2.2.1. Grupo 1: golpe delantero grave

En este primer grupo, que contiene el 8,21% de los siniestros analizados, podemos encontrar todos los golpes que afectan a la zona delantera del vehículo. Las piezas afectadas con más frecuencia se pueden observar en la siguiente tabla.

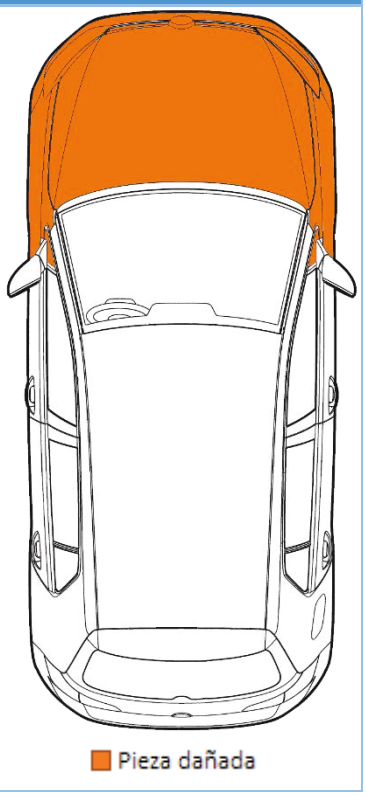
Variables significativas		Porcentaje	
63203	Paragolpes delantero (zona central)	99,83%	
50203	Componentes de la coraza (delanteros)	86,51%	
55103	Capó	83,23%	
94101	Faro delantero izquierdo	72,83%	
94102	Faro delantero derecho	70,33%	
66103	Luna custodia delantera	69,73%	
66101	Rejilla del radiador izquierda	66,69%	
66102	Rejilla del radiador derecha	66,65%	
66953	Anagrama del fabricante delantero	65,90%	
63201	Paragolpes delantero (zona izquierda)	62,67%	
63202	Paragolpes delantero (zona derecha)	62,61%	
50101	Aleta delantera izquierda	55,31%	
50102	Aleta delantera derecha	53,75%	
19103	Componentes del radiador	37,53%	
87103	Aire acondicionado central	34,68%	
19101	Componentes del radiador izquierda	31,79%	
19102	Componentes del radiador derecha	29,75%	
55101	Capó (zona izquierda)	20,66%	
55102	Capó (zona derecho)	20,44%	

Tabla 30: zona de impacto 1: golpe delantero grave

En este caso podemos observar como existen un total de 19 piezas que son afectadas con una frecuencia muy elevada. Podemos observar como suelen incluirse daños en las partes que se encuentran en la parte más adelantada del vehículo, como el paragolpes, los componentes de la coraza, el capó y los faros. Además podemos observar que algunas piezas superficiales que están situadas en zonas menos adelantadas también son dañadas, pero con menos frecuencia, como las aletas delanteras y las lunas custodia. Por último, también se puede observar que algunas piezas internas del motor del vehículo se suelen ver afectadas en este tipo de siniestros, como el radiador o el aire acondicionado.

Lo anteriormente mencionado nos demuestra que dentro de este grupo se encuentran todos los golpes delanteros que hayan sido de cierta gravedad, no afectando únicamente a piezas superficiales, sino también a piezas del interior del motor así como a las piezas laterales delanteras. Por tanto, este tipo de siniestro estará generalmente originado por un alcance o un golpe del vehículo contra otro o contra obstáculos frontalmente.

Existe una gran homogeneidad entre los siniestros, lo que demuestra que en la mayoría de las ocasiones no sólo será afectada una única pieza, sino que se verán afectadas varias, y siempre siguiendo un patrón similar.

Para continuar se realizará un análisis estadístico sobre los costes de mano de obra, pintura y sustitución de piezas de los siniestros de este clúster.

	Tot_mo *Coste de mano de obra*	Tot_pint *Coste de pintura de piezas*	Tot_sust *Coste de piezas sustituidas*
Mínimo	0,00 €	0,00 €	0,00 €
Máximo	5.063,80 €	2.432,80 €	26.565,00 €
Media	557,62 €	376,72 €	1.945,70 €
Cuartil 1	208,39 €	255,00 €	783,67 €
Mediana	350,86 €	342,26 €	1.274,70 €
Cuartil 3	692,60 €	451,69 €	2.317,70 €
Desv. Típica	545,66 €	202,76 €	2.008,30 €

Tabla 31: zona de impacto 1: distribución de costes

Como se puede observar en la tabla anterior, los siniestros contenidos en este clúster tienen unos costes medios, por lo general, altos o muy altos. Estos siniestros tienen unos costes de sustitución muy altos, estando la mediana en 1.274,70€, y el primer y tercer cuartil muy elevados. Los costes de reparación y pintura también son elevados, pero no tanto como los de sustitución de piezas. Esto nos demuestra que se trata de siniestros de carácter grave, en los que no ha bastado con la reparación de piezas sino que también ha sido necesario sustituir algunas de ellas. Adicionalmente, la desviación típica de los costes de mano de obra y, especialmente, de sustitución, son elevados, lo que quiere decir que serán dos variables que, además de altas (como se ha visto anteriormente con los cuartiles y la mediana), serán muy dispersas.

Este tipo de siniestros, por tanto, requerirán habitualmente sustitución de piezas y la mano de obra que esto conlleva, así como pequeños trabajos de pintura.

A continuación se ha realizado un análisis de correlaciones entre las variables que se están estudiando en este clúster, el cual nos indicará cómo están relacionadas entre sí estas variables.

	Tot_mo *Coste de mano de obra*	Tot_pint *Coste de pintura de piezas*	Tot_sust *Coste de piezas sustituidas*
Tot_mo *Coste de mano de obra*	1	0,7374	0,6730
Tot_pint *Coste de pintura de piezas*	0,7374	1	0,4731
Tot_sust *Coste de piezas sustituidas*	0,6730	0,4731	1

Tabla 32: correlación entre los costes en siniestros de la zona de impacto 1

En esta tabla se puede observar que existe una correlación elevada entre el coste de mano de obra y el resto de las variables. Esta relación es además positiva, por lo que se espera que una subida en el coste de mano de obra repercuta en el resto de las variables así como que una subida en cualquier variable repercuta en el coste de mano de obra también de forma positiva. Los costes de pintura y de sustitución de piezas también tendrán una correlación positiva, pero en este caso en una medida moderada.

A continuación se ha realizado la separación en clústeres de los siniestros que corresponden a esta zona de impacto. Se han obtenido dos grupos claramente definidos, y la distribución de los siniestros entre los mismos se puede observar en la siguiente ilustración.

Ilustración 33: distribución de la zona de impacto 1 por severidad

Como se puede observar, existen dos grupos, donde el primero de ellos contiene el 41,30% de los datos, mientras que el segundo de ellos contiene el resto: 58,70% de los datos. Para continuar se han realizado ciertos análisis estadísticos sobre los siniestros de cada uno de estos clústeres, los cuales han sido recogidos en las tablas que se pueden observar a continuación.

Clúster	Variable	Mínimo	Máximo	Media	Desv. Típica
1: Severidad alta	*Tot_mo*	4,71 €	3.699,60 €	747,47 €	464,63 €
	Tot_pint	0,00 €	1.766,80 €	445,96 €	192,46 €
	Tot_piez	1.153,80 €	26.565,00 €	2.461,70 €	958,85 €
2: Severidad media con gasto elevado en piezas	*Tot_mo*	0,00 €	2.105,20 €	259,85 €	150,88 €
	Tot_pint	0,00 €	1.772,00 €	288,64 €	124,16 €
	Tot_piez	0,00 €	1.502,00 €	818,15 €	308,95 €

Tabla 33: análisis sobre los grupos de severidad zona de impacto 1 (I)

Clúster	Variable	Q1	Mediana	Q3
1: Severidad alta	*Tot_mo*	407,70 €	615,26 €	962,62 €
	Tot_pint	332,58 €	414,39 €	529,59 €
	Tot_piez	1.700,90 €	2.181,00 €	3.033,30 €
2: Severidad media con gasto elevado en piezas	*Tot_mo*	158,45 €	224,11 €	319,41 €
	Tot_pint	215,12 €	285,36 €	353,69 €
	Tot_piez	587,48 €	817,31 €	1.054,90 €

Tabla 34: análisis sobre los grupos de severidad zona de impacto 1 (II)

Como se puede observar en las tablas anteriores, el primer grupo de severidad muestra unos valores muy elevados en las tres variables. En primer lugar, en el coste de mano de obra, a pesar de su mínimo reducido, el primer cuartil tiene un valor elevado, lo que significa que el 75% de los siniestros de este grupo tienen un coste de pintura superior a 407,70€. Además, se puede observar que la desviación típica de esta variable es también alta (464,63€), lo que significa que esta variable es bastante dispersa. Con respecto al coste de pintura, se puede observar algo muy similar al coste de mano de obra: se observan unos cuartiles con valores altos, donde el 75% de los siniestros tiene un coste superior a 332,58€. Sin embargo, en este caso la desviación típica es media, por lo que se espera que esta variable sea mucho más estable. Por último, el coste de sustitución de piezas no sólo tiene una media muy elevada (2.461,70€), sino que también los valores de los cuartiles avalan los elevadísimos costes de este tipo que generan estos siniestros. De esta forma, se puede afirmar que el 75% de estos siniestros tienen un coste de sustitución de piezas superior a 1.700,90€. Adicionalmente, la desviación típica de esta variable es muy elevada, lo cual además demuestra que tendrá una dispersión muy alta.

Por otra parte, el segundo grupo de siniestros muestra unos valores, por lo general, más reducidos que el primer grupo. En este grupo se puede observar como Los costes de mano de obra y pintura se comportan de una forma similar, teniendo costes de entre 150 y 360 euros en sus cuartiles y mediana, y también tienen una desviación típica baja, lo cual indica que no sólo son de severidad media sino que además tienen poca dispersión. Por otra parte, el coste de sustitución de piezas tiene unos cuartiles mucho más elevados, lo cual nos es demostrado porque el 50% de los siniestros de este tipo tendrán un coste comprendido en el intervalo [587,48, 1054,90]. El otro 50% se reparte a partes iguales entre los valores inferiores y superiores a dicho intervalo. La dispersión, en este caso, es media, de manera que no habrá una dispersión excesiva en los costes de sustitución de piezas de este tipo de siniestros.

A continuación se puede observar la dispersión de estos clústeres representados en un gráfico en tres dimensiones. En este caso, se ha decidido que todos los ejes deben mostrar el rango [0, 4.000] para apreciar claramente la dispersión de estos siniestros, ya que a partir de dicha cantidad hay muy pocas observaciones y dificultarían la comprensión de este gráfico.

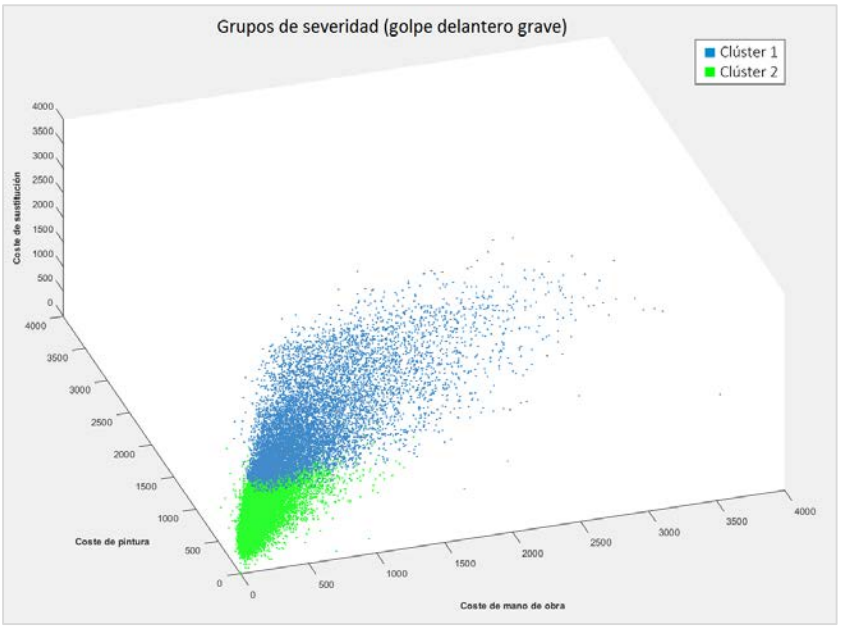


Ilustración 34: gráfico 3D de la distribución de la zona de impacto 1 por severidad

En este gráfico se puede observar claramente los datos sobre la dispersión y la formación de estos dos clústeres. En concreto, se puede ver como el primer clúster toma valores altos, especialmente en mano de obra y sustitución de piezas y es muy disperso, especialmente en el coste de sustitución de piezas. En el caso del segundo clúster, se puede ver como es mucho menos disperso.

Como se ha podido ver, el primer clúster se corresponde con siniestros de severidad alta, de acuerdo a la clasificación que se hizo en el apartado 4.1, y el segundo se corresponde con siniestros de severidad media con un gasto elevado en sustitución de piezas.

4.2.2.2. Grupo 2: golpe delantero leve

En este segundo grupo, que contiene el 34,32% de los siniestros, podemos encontrar también golpes frontales en el vehículo, pero se observan ligeras diferencias con respecto al grupo anterior.

Para comenzar, se ha recogido en una tabla la información sobre las órdenes de reparación, pintura y sustitución de piezas más frecuentes en este tipo de siniestros.

	Variables significativas	Porcentaje
63203	Paragolpes delantera central	41,13%
64103	Parabrisas	16,51%
50101	Aleta delantera izquierda	13,40%
50102	Aleta delantera derecha	12,27%
55103	Capó	12,15%
66953	Anagrama del fabricante delantero	10,22%
94101	Faro delantero izquierdo	6,54%
63202	Paragolpes delantero (zona derecha)	6,25%
63201	Paragolpes delantero (zona izquierda)	5,95%
94102	Faro delantero derecho	5,76%

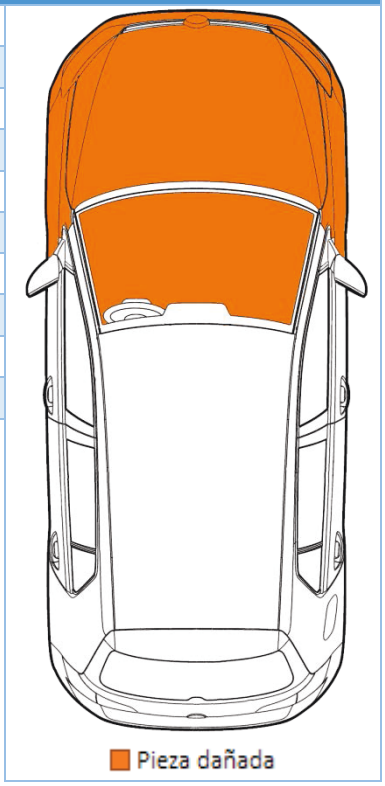


Tabla 35: zona de impacto 2: golpe delantero leve

En este caso se puede observar como existe una similitud muy elevada en las piezas afectadas con respecto al primer grupo, pero sin embargo vemos como no aparecen elementos internos del motor del vehículo, y a cambio aparece el parabrisas. Sin embargo, podemos observar como este grupo es muy heterogéneo, con apenas una única pieza que aparece con una frecuencia significativa (paragolpes delantero).

Todo esto demuestra que este grupo está más vagamente definido, pero sobre todo que es un grupo en el que los siniestros suelen tener un número mucho menor de piezas afectadas en cada siniestro. Estos golpes serán, por lo general, más leves, ya que las piezas afectadas son mayoritariamente superficiales, y no existen piezas del motor afectadas con frecuencia.

Por lo general, se espera que este tipo de siniestros estén ocasionados por roces con otros vehículos u objetos o actos de vandalismo.

A continuación, al igual que en los casos anteriores, se analizarán diversos datos estadísticos sobre los costes de mano de obra, pintura y sustitución de piezas sobre los siniestros contenidos en este grupo.

	Tot_mo *Coste de mano de obra*	Tot_pint *Coste de pintura de piezas*	Tot_sust *Coste de piezas sustituidas*
Mínimo	0,00 €	0,00 €	0,00 €
Máximo	24.800,00 €	3.000,00 €	29.632,00 €
Media	107,08 €	110,08 €	266,22 €
Cuartil 1	41,04 €	0,00 €	9,00 €
Mediana	67,15 €	91,48 €	126,22 €
Cuartil 3	113,00 €	173,71 €	214,20 €
Desv. Típica	221,10 €	133,59 €	893,05 €

Tabla 36: zona de impacto 2: distribución de costes

A través de esta tabla se pueden obtener diversas conclusiones de interés para el estudio que se está llevando a cabo. En primer lugar, se puede observar como las tres variables tienen medianas bajas, demostrando que en el 50% de los casos estos siniestros implican un coste económico bajo o muy bajo.

Con respecto al coste de mano de obra se puede afirmar que el 75% de estos siniestros tendrán un coste inferior o igual a 113,00€, dado que este es el valor del tercer cuartil. Adicionalmente, se puede afirmar que esta variable tiene una desviación típica media-baja, y dado el valor tan pequeño que tienen los valores de los cuartiles, se espera que los siniestros dentro del 75% más bajo sean menos dispersos, mientras que el resto será bastante más disperso.

Por otra parte, con respecto al coste de pintura se puede afirmar que en al menos el 25% de los casos no existirá este coste. Además, en el 50% de los siniestros estudiados tiene unos costes de pintura inferiores o iguales a 91,48€, como nos indica el valor de la mediana. Además, la desviación típica de esta variable es la más reducida de las tres, lo que quiere decir que es la variable más concentrada de las tres.

Por último, el coste de piezas sustituidas es la variable cuyos valores son más elevados de las tres analizadas, pero no de una forma muy grande. En este caso, el primer cuartil nos indica que el 25% de los siniestros tendrán un coste de 9,00€ o menos. Por otra parte, el 75% de los siniestros tienen un gasto en sustitución de piezas inferior o igual a 214,20€, lo cual se ve reflejado a través del tercer cuartil. Sin embargo, el valor de la desviación típica es muy elevado (893,05€), lo cual indica que los datos son muy dispersos, especialmente en los valores más altos (a partir del tercer cuartil).

A continuación se ha realizado un análisis de correlaciones entre las variables que se están estudiando en este clúster, el cual nos indicará cómo están relacionadas entre sí estas variables.

	Tot_mo *Coste de mano de obra*	Tot_pint *Coste de pintura de piezas*	Tot_sust *Coste de piezas sustituidas*
Tot_mo *Coste de mano de obra*	1	0,2889	0,3345
Tot_pint *Coste de pintura de piezas*	0,2889	1	0,0945
Tot_sust *Coste de piezas sustituidas*	0,3345	0,0945	1

Tabla 37: correlación entre los costes en siniestros de la zona de impacto 2

En esta tabla se puede observar que existe una correlación bastante reducida entre todas las variables. En el caso del coste de mano de obra, su correlación con las demás variables es reducida y positiva, de manera que ante una subida en el coste de mano de obra, se espera que también se produzca una subida en el resto de las variables, pero de una forma más moderada. Lo mismo ocurre si se produce una subida en cualquier otra variable: se espera que repercuta en el coste de mano de obra también de forma positiva y moderada. Los costes de pintura y de sustitución de piezas también tendrán una correlación positiva, pero en este caso en una medida muy reducida, apenas 0,0945.

Todo lo mencionado anteriormente se puede observar claramente en el gráfico de dispersión en 3 dimensiones que se ha realizado. En este gráfico se ha utilizado una escala [0,2.000] en los tres ejes para asegurar que se aprecien bien las observaciones así como su dispersión en datos elevados. Existen muy pocas observaciones fuera de dicha escala, y su inclusión dificultaría mucho la observación de donde se encuentran la mayor parte de los datos.

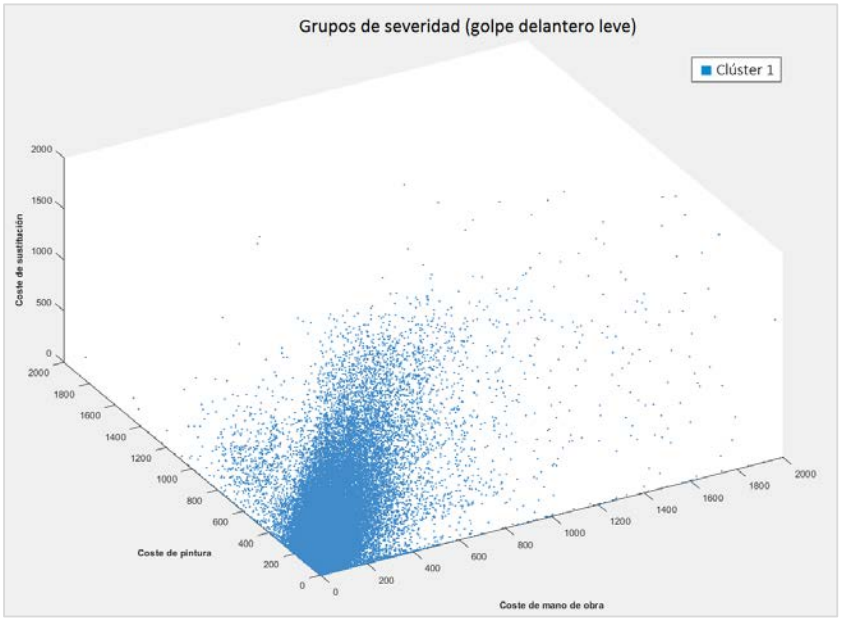


Ilustración 35: gráfico 3D de la distribución de la zona de impacto 2 por severidad

En este gráfico se puede observar como todas las observaciones se agrupan en torno a valores muy bajos, tendiendo a cero, mientras que cuando se van alejando de éstos son cada vez más y más dispersos.

Por todo lo anteriormente detallado, se puede esperar que esta zona de impacto tenga con frecuencia costes bajos o muy bajos, pero también existen observaciones con costes elevados.

En este caso se ha intentado realizar la subdivisión del grupo en distintos grupos de severidad, pero esto no ha sido posible debido a que los grupos obtenidos no mostraban características comunes. El algoritmo de agrupamiento simplemente agrupa los datos en clústeres de un tamaño similar entre sí, los cuales una vez representados en un gráfico en tres dimensiones, formaban cortes en línea recta de la base de datos original, lo cual nos demuestra que el algoritmo ha sido incapaz de encontrar grupos con características específicas sino que simplemente hace una división, de alguna forma, a partes iguales. Sin embargo, podríamos enmarcar estos siniestros en el grupo de siniestros de severidad baja, de acuerdo a lo estudiado en el apartado 4.1.

4.2.2.3. Grupo 3: golpe derecho

En este clúster, que representa el 13,51% de los siniestros, se agrupan todos aquellos siniestros cuya zona de impacto es la parte derecha del vehículo. En la siguiente tabla se pueden observar las zonas de impacto más frecuentes en este grupo.

	Variables significativas	Porcentaje
66202	Molduras derecha	90,88%
57102	Puerta delantera derecha	82,30%
53102	Aleta trasera derecha	69,25%
50102	Aleta delantera derecha	56,19%
58102	Puerta trasera derecha	46,83%
63303	Paragolpes trasero (zona central)	29,05%
63203	Paragolpes delantero (zona central)	25,07%
66502	Retrovisor derecho	19,89%
51752	Estribo derecho	12,15%
57202	Mecanismo de cierre de puerta delantero derecho	7,35%

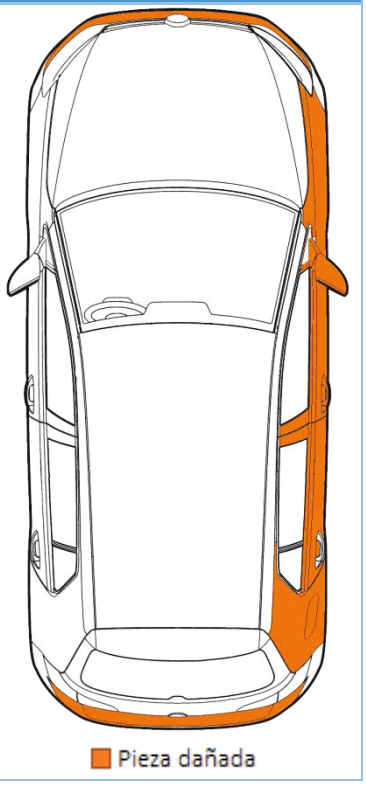


Tabla 38: zona de impacto 3: golpe derecho

Como se puede observar, es muy habitual en este grupo que resulten dañadas las molduras de la zona derecha, las puertas derechas y las aletas de este mismo lado del vehículo. Además, con menos frecuencia aparecen los paragolpes delantero y trasero, así como el retrovisor derecho, el estribo y el mecanismo de cierre de puerta de este mismo lado del vehículo.

Este grupo muestra una homogeneidad alta, de forma que se puede esperar que la mayoría de los siniestros de este tipo tengan dañados, al menos, las molduras, la puerta delantera y las aletas, todo ello de la zona derecha vehículo. También se puede esperar que en los siniestros de este grupo sean dañadas varias piezas al mismo tiempo, por lo que no se tratará de siniestros leves.

Por tanto, este tipo de siniestros estará originado, por ejemplo, por rozaduras laterales con otros vehículos, paredes, columnas u otros objetos por la parte derecha del vehículo.

A partir de este clúster, se ha generado una tabla que contiene datos estadísticos sobre los costes de mano de obra, pintura y sustitución de piezas de los siniestros contenidos en este grupo, de manera que se puedan conocer más a fondo las características de los siniestros que pertenecen al mismo.

	Tot_mo *Coste de mano de obra*	Tot_pint *Coste de pintura de piezas*	Tot_sust *Coste de piezas sustituidas*
Mínimo	0,00 €	0,00 €	0,00 €
Máximo	4.735,50 €	1.961,50 €	25.686,00 €
Media	271,02 €	345,39 €	253,23 €
Cuartil 1	114,40 €	229,97 €	8,77 €
Mediana	192,08 €	307,94 €	20,42 €
Cuartil 3	320,00 €	419,75 €	270,27 €
Desv. Típica	283,78 €	173,34 €	760,05 €

Tabla 39: zona de impacto 3: distribución de costes

Como se puede observar, se trata por lo general de siniestros cuyo coste de pintura es bastante elevado con respecto a los costes de mano de obra y de sustitución de piezas, como se puede observar en los valores de la media, la mediana y los cuartiles. Este coste tiene una desviación típica bastante reducida, lo que indica que es bastante estable.

Por otra parte, los costes de sustitución son generalmente bajos, ya que la mediana está en apenas 20,42€. A pesar de esto, esta variable tiene una media de 253,23€, la cual es muy superior a la mediana, y junto con el valor elevado de su desviación típica demuestra que es una variable que será dispersa en los valores altos y concentrada en los valores bajos.

Por otra parte, el coste de mano de obra tiene un valor medio comparado con las otras dos variables. Además, su media, mediana y cuartiles se encuentran en posiciones muy cercanas entre

ellas, lo que junto con la reducida desviación típica demuestra que es una variable que no es muy dispersa.

A continuación se ha realizado un análisis de correlaciones entre las variables que se están estudiando en este clúster, el cual nos indicará cómo están relacionadas entre sí estas variables.

	Tot_mo *Coste de mano de obra*	Tot_pint *Coste de pintura de piezas*	Tot_sust *Coste de piezas sustituidas*
Tot_mo *Coste de mano de obra*	1	0,5436	0,5249
Tot_pint *Coste de pintura de piezas*	0,5436	1	0,2177
Tot_sust *Coste de piezas sustituidas*	0,5249	0,2177	1

Tabla 40: correlación entre los costes en siniestros de la zona de impacto 3

Al igual que en los casos anteriores, existe una correlación moderada y positiva entre el coste de mano de obra con las demás variables, de manera que ante una subida en el coste de mano de obra, se espera que también se produzca una subida en el resto de las variables, pero de una forma moderada. Lo mismo ocurre si se produce una subida en cualquier otra variable: se espera que repercuta en el coste de mano de obra también de forma positiva y moderada. Los costes de pintura y de sustitución de piezas también tendrán una correlación positiva, pero en este caso en una medida más reducida.

Tras la ejecución del algoritmo de clustering sobre estos datos se han obtenido un total de tres grupos de severidad con características diferentes entre sí. La distribución de los golpes derechos en grupos de severidad se puede observar en la siguiente ilustración.

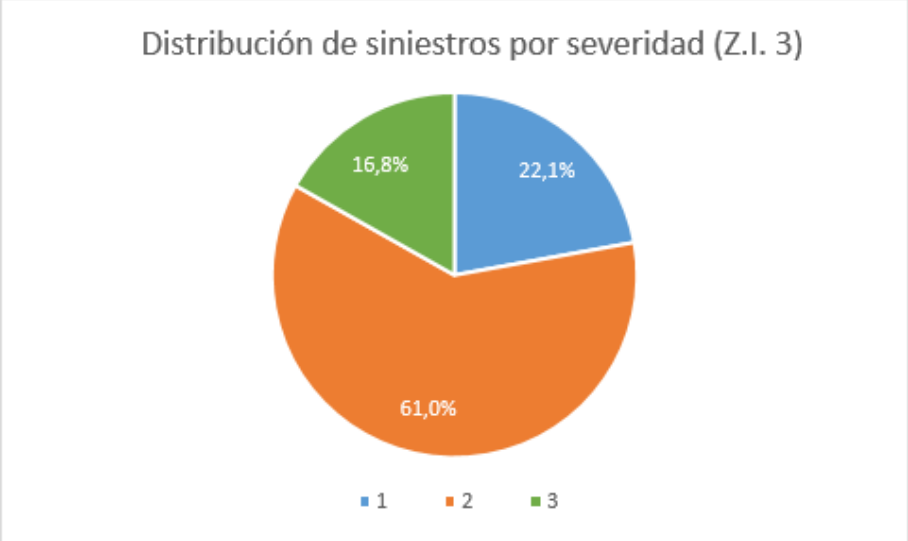


Ilustración 36: distribución de la zona de impacto 3 por severidad

Como se puede observar, existen tres grupos de severidad, entre los cuales el segundo contiene la mayoría de los siniestros (61,03%). Los otros dos grupos contienen el resto de los siniestros y,

a pesar de su reducido tamaño, presentan características que los definen a la perfección. En las siguientes tablas se pueden observar algunos datos estadísticos sobre estos tres clústeres.

Clúster	Variable	Mínimo	Máximo	Media	Desv. Típica
1: Severidad media con gasto elevado en piezas	*Tot_mo*	26,40 €	2.081,50 €	369,40 €	207,28 €
	Tot_pint	0,00 €	1.039,10 €	381,47 €	118,53 €
	Tot_piez	8,28 €	25.686,00 €	449,02 €	339,94 €
2: Severidad baja	*Tot_mo*	0,00 €	629,65 €	144,88 €	80,34 €
	Tot_pint	0,00 €	497,30 €	251,96 €	78,31 €
	Tot_piez	0,00 €	416,53 €	37,70 €	69,16 €
3: Severidad media con gasto elevado en pintura	*Tot_mo*	0,00 €	1.066,50 €	324,33 €	146,11 €
	Tot_pint	273,62 €	1.350,50 €	556,84 €	170,55 €
	Tot_piez	0,00 €	473,14 €	47,80 €	64,01 €

Tabla 41: análisis sobre los grupos de severidad zona de impacto 3 (I)

Clúster	Variable	Q1	Mediana	Q3
1: Severidad media con gasto elevado en piezas	*Tot_mo*	212,58 €	319,20 €	488,26 €
	Tot_pint	298,58 €	366,31 €	448,17 €
	Tot_piez	295,38 €	396,33 €	577,56 €
2: Severidad baja	*Tot_mo*	83,70 €	132,30 €	191,74 €
	Tot_pint	194,46 €	248,35 €	306,22 €
	Tot_piez	6,12 €	9,61 €	23,38 €
3: Severidad media con gasto elevado en pintura	*Tot_mo*	222,95 €	303,48 €	402,55 €
	Tot_pint	442,32 €	507,53 €	619,45 €
	Tot_piez	9,62 €	19,60 €	51,90 €

Tabla 42: análisis sobre los grupos de severidad zona de impacto 3 (II)

Para comenzar, podemos encontrar en el primer clúster unos costes de mano de obra y pintura medios, lo cual es demostrado a través de sus cuartiles 1 y 3 y su mediana, los cuales contienen valores medios. Además, la desviación típica de estas variables es bastante reducida, lo que indica que no tienen mucha dispersión. Sin embargo, el coste de sustitución de piezas tiene también un coste medio, pero más elevado por lo general, y con una desviación típica mucho más elevada, la cual indica que el coste de sustitución es mucho más disperso y, por tanto, variable entre unos siniestros y otros. Serán, por tanto, siniestros leves pero con gastos más elevados en sustitución de piezas.

Por otra parte, el segundo clúster muestra costes de pintura, sustitución de piezas y mano de obra por lo general muy bajos. En este caso los costes de mano de obra son muy reducidos, con el 75% de los siniestros por debajo de 191,74€ y una desviación típica de apenas 80,34€. Con respecto a los gastos de pintura, también son bastante bajos, ya que el 75% de estos siniestros tienen un coste de pintura inferior a 306,22€. Sin embargo, los gastos de pintura son los más elevados en este clúster, aunque no toman un valor muy elevado. Por otra parte, los gastos de sustitución de

piezas son en el 75% de los casos inferiores a 51,90€, por lo que se espera que estos gastos se deban a sustitución de piezas pequeñas como molduras. Por todo esto, podemos afirmar que se trata de siniestros de severidad baja.

Por último, el tercer clúster muestra unas características similares en cierta medida al tercero de ellos, sólo que muestra unos costes muy reducidos de sustitución de piezas, y unos mucho más elevados de pintura. Adicionalmente, se pueden observar desviaciones típicas muy reducidas en las tres variables, por lo que este clúster tiene una dispersión muy baja. De esta forma, se entiende que se trata de siniestros de severidad leve con gastos elevados en trabajos de pintura.

Como se puede apreciar, los clústeres muestran una similitud bastante alta con los definidos al realizar el análisis de severidad en el apartado 4.1. En este caso podemos apreciar siniestros de severidad baja, cuyo rango y costes medios son más bajos incluso que en el caso de los sinestros de severidad baja definidos anteriormente, pero es comprensible dado que estamos tratando un tipo específico de siniestros y no la totalidad de los datos, por lo que es normal que no tengan exactamente la misma forma. Por otra parte, también se pueden encontrar los dos tipos de siniestros de severidad media que se apreciaban en el análisis inicial (con gasto elevado en piezas y con gasto elevado en pintura).

Adicionalmente, se puede afirmar que los grupos han sido construidos correctamente dado que la desviación típica que muestran todos ellos es bastante reducida, lo cual indica que los siniestros de cada grupo son bastante similares entre sí.

A continuación se pueden observar la dispersión de estos grupos de siniestros representados en un gráfico en tres dimensiones.

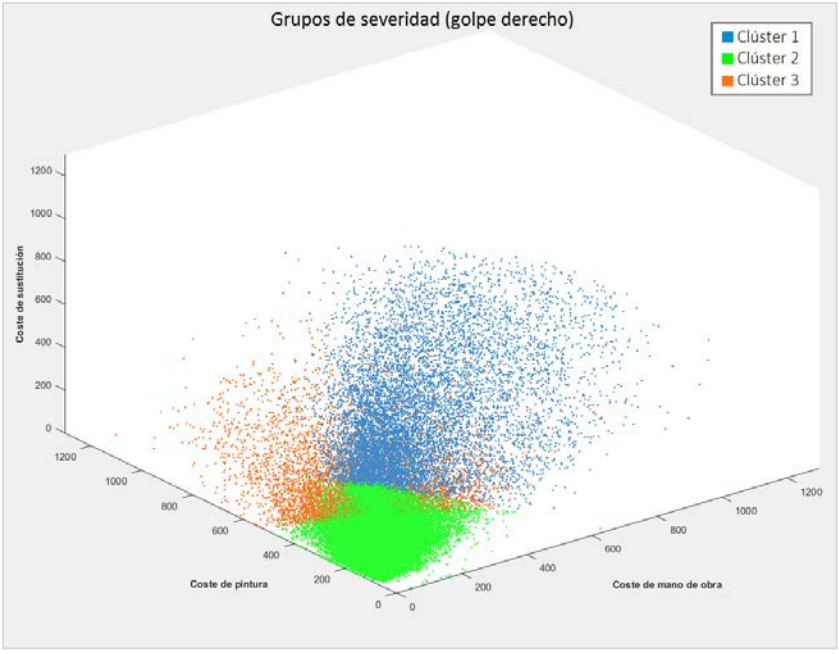


Ilustración 37: gráfico 3D de la distribución de la zona de impacto 3 por severidad

En este gráfico se pueden apreciar claramente los tres grupos de severidad analizados así como las características de los mismos que se han comentado con anterioridad. En primer lugar, el clúster número 1 y el número 3 son un poco más dispersos que el segundo clúster. También se puede observar que el clúster número 2 es el de más bajos costes, así como que el grupo 1 tiene unos costes de sustitución más elevados, y el tercer grupo, en cambio, tiene unos costes de pintura más elevados. Los tres clúster tienen la mayor parte de las observaciones en el rango [0,1000] en las tres variables, por lo que también se puede determinar que se trata de siniestros de severidad media-baja.

Por lo anteriormente comentado, podemos afirmar que se trata de siniestros de severidad media-baja, los cuales se corresponden con golpes laterales derechos.

4.2.2.4. Grupo 4: golpe izquierdo

En este grupo se pueden observar los golpes laterales que afectan al lado izquierdo del vehículo, los cuales suponen el 12,73% del total de los siniestros registrados en la base de datos. A continuación se ha elaborado una tabla que contiene las piezas dañadas más frecuentemente en este tipo de siniestros.

	Variables significativas	Porcentaje
57101	Puerta delantera izquierda	88,83%
66201	Molduras izquierda	84,30%
50101	Aleta delantera izquierda	61,63%
53101	Aleta trasera izquierda	53,88%
58101	Puerta trasera izquierda	38,63%
63203	Paragolpes delantero (zona central)	26,68%
66501	Retrovisor izquierdo	20,20%
63303	Paragolpes trasero (zona central)	18,90%
51751	Estribo izquierdo	9,95%
57201	Mecanismo de cierre de puerta delantero izquierdo	7,04%

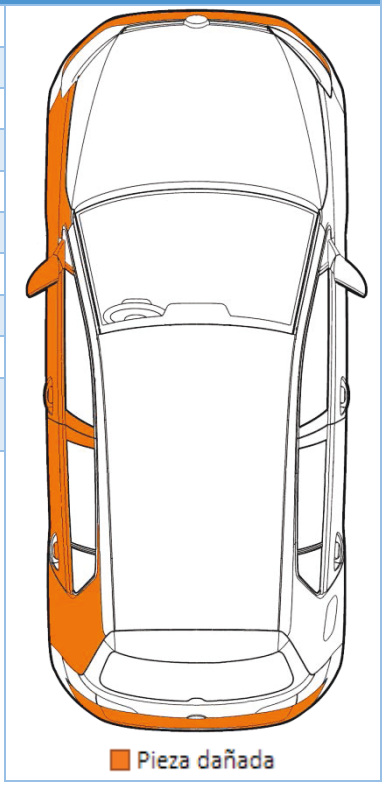


Tabla 43: zona de impacto 4: golpe izquierdo

Como se puede observar, este grupo es casi idéntico al grupo descrito con anterioridad. En este caso, también son afectadas con frecuencia las molduras, las puertas y las aletas situadas en el lateral izquierdo del vehículo.

En este caso también se observa una homogeneidad alta entre los siniestros del grupo, dado que existen varias piezas que resultan dañadas con una frecuencia muy elevada. De la misma manera que en el anterior grupo, se trata de golpes de gravedad media o alta, en los que se requiere la reparación de varias piezas de la zona izquierda del vehículo.

Por lo anteriormente mencionado, este tipo de siniestros también estará originado, por ejemplo, por rozaduras laterales con otros vehículos, paredes, columnas u otros objetos por la parte izquierda del vehículo.

Se ha elaborado una tabla que contiene información estadística sobre los costes que acarrean los siniestros contenidos en este clúster.

	Tot_mo *Coste de mano de obra*	Tot_pint *Coste de pintura de piezas*	Tot_sust *Coste de piezas sustituidas*
Mínimo	0,00 €	0,00 €	0,00 €
Máximo	4.626,40 €	1.617,60 €	38.519,00 €
Media	250,60 €	310,42 €	303,74 €
Cuartil 1	96,00 €	205,97 €	8,97 €
Mediana	163,24 €	277,96 €	30,97 €
Cuartil 3	281,60 €	380,06 €	301,66 €
Desv. Típica	299,43 €	152,04 €	961,53 €

Tabla 44: zona de impacto 4: distribución de costes

Como se puede observar, en este tipo de siniestro los costes de mano de obra son generalmente bajos, ya que el 75% de los siniestros suponen un coste inferior a 281,60€ en concepto de mano de obra. La desviación típica de esta variable es media, lo que quiere decir que es bastante dispersa y variará bastante de unos siniestros a otros. Por otra parte, el coste de pintura de piezas de los siniestros es un poco más elevado, teniendo un valor medio de 310,42€, y su mediana no está muy alejada de este valor. Su reducida desviación típica confirma que estos valores no son muy dispersos. Por último el coste de piezas sustituidas es por lo general bajo, ya que el 75% de estos siniestros tendrán un coste de sustitución de piezas de menos de 301,66€ y, de hecho, en el 50% de los siniestros es inferior a 30,97€. Sin embargo, la desviación típica de esta variable es muy elevada (961,53€), por lo que se espera que los valores altos sean muy dispersos.

A continuación se ha realizado un análisis de correlaciones entre las variables que se están estudiando en este clúster, el cual nos indicará cómo están relacionadas entre sí estas variables.

	Tot_mo *Coste de mano de obra*	Tot_pint *Coste de pintura de piezas*	Tot_sust *Coste de piezas sustituidas*
Tot_mo *Coste de mano de obra*	1	0,5965	0,5044
Tot_pint *Coste de pintura de piezas*	0,5965	1	0,2688
Tot_sust *Coste de piezas sustituidas*	0,5044	0,2688	1

Tabla 45: correlación entre los costes en siniestros de la zona de impacto 4

Al igual que en los casos anteriores, existe una correlación moderada y positiva entre el coste de mano de obra con las demás variables, de manera que ante una subida en el coste de mano de obra, se espera que también se produzca una subida en el resto de las variables, pero de una forma moderada. Lo mismo ocurre si se produce una subida en cualquier otra variable: se espera que repercuta en el coste de mano de obra también de forma positiva y moderada. Los costes de pintura y de sustitución de piezas también tendrán una correlación positiva, pero en este caso en una medida más reducida.

Se ha llevado a cabo el proceso de agrupamiento sobre los siniestros contenidos en este clúster, y se han obtenido un total de 4 grupos, entre los cuales se reparten todos los siniestros. La distribución de estos siniestros en los grupos obtenidos se puede observar en el gráfico expuesto a continuación.

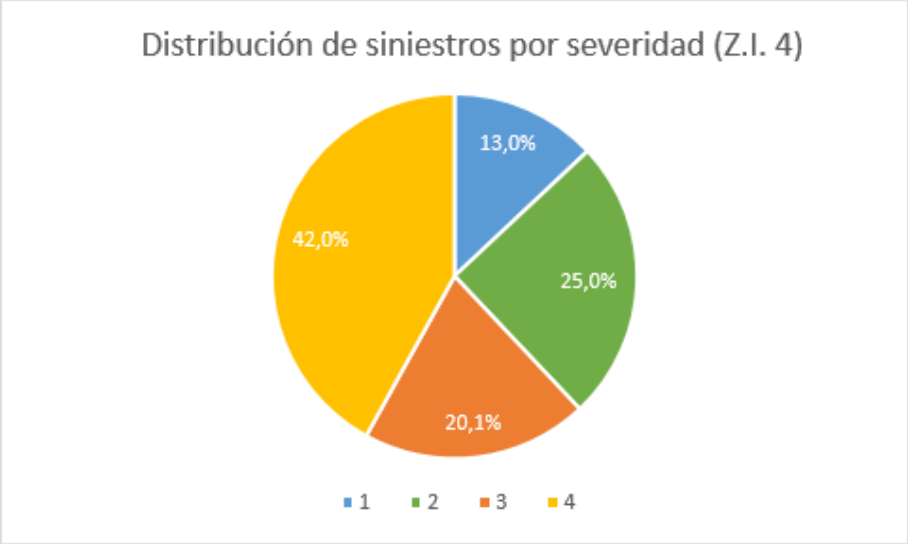


Ilustración 38: distribución de la zona de impacto 4 por severidad

Como se puede observar, existen cuatro grupos entre los que se reparten todos los siniestros cuya zona de impacto es lateral izquierdo. A continuación se han realizado ciertos análisis estadísticos sobre estos clústeres para comprender mejor su formación y características.

Clúster	Variable	Mínimo	Máximo	Media	Desv. Típica
1: Severidad media con pintura y mano de obra	*Tot_mo*	0,00 €	1.213,70 €	390,86 €	171,34 €
	Tot_pint	123,95 €	1.210,30 €	496,65 €	150,22 €
	Tot_piez	0,00 €	621,64 €	111,78 €	117,12 €
2: Severidad baja con gasto en pintura	*Tot_mo*	0,00 €	288,60 €	130,93 €	61,28 €
	Tot_pint	211,91 €	768,86 €	333,60 €	72,20 €
	Tot_piez	0,00 €	321,14 €	36,00 €	57,93 €
3: Severidad media	*Tot_mo*	32,97 €	2.645,50 €	292,59 €	191,53 €
	Tot_pint	0,00 €	1.142,00 €	331,48 €	109,75 €
	Tot_piez	244,81 €	38.519,00 €	520,03 €	475,43 €
4: Severidad baja	*Tot_mo*	0,00 €	613,60 €	130,08 €	77,46 €
	Tot_pint	0,00 €	327,36 €	193,25 €	52,38 €
	Tot_piez	0,00 €	357,45 €	50,45 €	79,21 €

Tabla 46: análisis sobre los grupos de severidad zona de impacto 4 (I)

Clúster	Variable	Q1	Mediana	Q3
1: Severidad media con pintura y mano de obra	*Tot_mo*	277,41 €	350,86 €	460,07 €
	Tot_pint	392,54 €	474,78 €	577,09 €
	Tot_piez	14,40 €	52,16 €	188,36 €
2: Severidad baja con gasto en pintura	*Tot_mo*	86,22 €	127,60 €	172,80 €
	Tot_pint	280,49 €	319,94 €	372,87 €
	Tot_piez	8,38 €	9,78 €	29,38 €
3: Severidad media	*Tot_mo*	151,20 €	240,00 €	385,80 €
	Tot_pint	253,59 €	314,00 €	393,85 €
	Tot_piez	338,06 €	464,88 €	634,65 €
4: Severidad baja	*Tot_mo*	70,76 €	113,06 €	175,28 €
	Tot_pint	156,93 €	192,94 €	230,35 €
	Tot_piez	6,01 €	9,43 €	59,46 €

Tabla 47: análisis sobre los grupos de severidad zona de impacto 4 (II)

En el primero de estos grupos se puede observar como los costes de mano de obra toman un valor medio, siendo en la mitad de los casos inferiores a 350,86€. Por otra parte, el coste de pintura es un poco más elevado, partiendo de un mínimo de 123,95€ y rondando valores aproximadamente 100€ por encima de los costes de mano de obra. Por otra parte, los gastos de sustitución de piezas toma valores más reducidos, siendo en el 50% de los casos inferiores a 52,16€. Además, el máximo de esta variable son apenas 621,24€, lo que reafirma que se trata de una variable que tomará valores reducidos. Por último, si se observa la desviación típica de estas tres variables se puede ver que son bastante reducidos, por lo que este grupo no estará muy disperso. Por tanto, se puede afirmar que se tratará de siniestros de severidad media con gastos medios en pintura y mano de obra.

Con respecto al segundo clúster, los costes de mano de obra se mueven en un rango muy reducido: [0, 288,60]. En este clúster, el 75% de los siniestros tendrán un coste de mano de obra inferior a

172,80€, lo cual reafirma que estos siniestros tendrán un coste bajo de este tipo. Adicionalmente, los gastos de sustitución de piezas son todavía más reducidos a pesar de que se mueven en un rango mayor ([0, 321,14€]), ya que el 75% de los siniestros tienen costes de sustitución de piezas inferiores a 29,38€. Con respecto al coste de pintura, se puede ver que es medio, y de hecho no sólo tiene una media de 333,60€ sino que el 50% de los siniestros tienen costes de este tipo de entre 280,49€ y 372,87€. Por último, la desviación típica de estas tres variables es muy reducida, lo que quiere decir que el grupo entero será muy poco disperso. Dado que la mayor parte de los costes en los que se incurre en este clúster son de pintura, se espera que sean siniestros leves ocasionados por rozaduras leves, que sólo requieran este tipo de trabajos.

El tercer clúster encontrado contiene siniestros de una envergadura un poco mayor. En este caso, los gastos de pintura y mano de obra toman valores medios, teniendo medias y cuartiles de cierta similitud. De hecho, estas dos variables tienen una desviación típica reducida, por lo que se puede comprobar cómo estas variables se mantendrán en valores medios por lo general. Sin embargo, los gastos en sustitución de piezas son un poco más elevados, ya que el 50% de los siniestros tendrán un coste de sustitución de piezas superior a 464,88€. Además, la desviación típica de esta variable es bastante más elevada que la de las otras dos, y se espera que dé lugar a valores más dispersos. Estos siniestros, por tanto, serán siniestros de severidad media.

Por último, el cuarto clúster es el que tiene los valores más bajos de los cuatro que han resultado de este análisis. El coste más reducido de todos es el coste de sustitución de piezas, donde la media es de 50,45€ y el 75% de los siniestros toma un valor inferior a 59,46€. El coste de mano de obra también es muy reducido, ya que el 75% de los siniestros tienen un coste de este tipo inferior a 175,28€. Por último, el coste de pintura es el más elevado de los tres, pero aun así el 75% de los siniestros de este tipo tienen un coste de este tipo inferior a 230,35€. Adicionalmente, la desviación típica de las tres variables es muy reducida, lo que indica que él clúster será muy poco disperso. Este tipo de siniestro, por tanto, se podrán identificar como de severidad baja.

Si se representan los datos de estos cuatro grupos sobre un gráfico de dispersión de tres dimensiones se puede observar claramente las características detalladas con anterioridad. En este caso se ha reducido la escala a [0, 1.200] en los tres ejes dado que la cantidad de observaciones que hay por encima de dichos valores es muy reducida y dificulta la correcta visualización de los grupos formados.

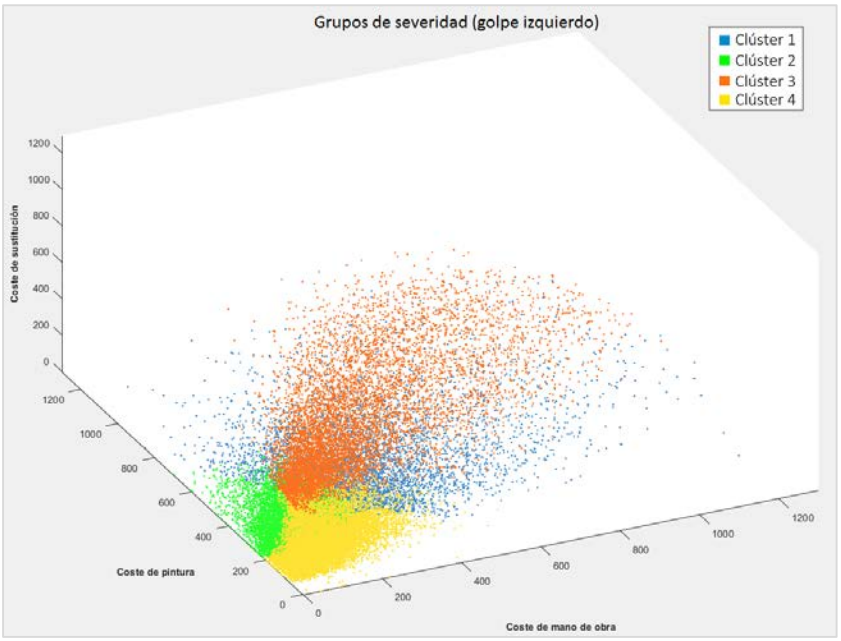


Ilustración 39: gráfico 3D de la distribución de la zona de impacto 4 por severidad

En el caso de esta zona de impacto, se pueden encontrar cuatro subgrupos que también se pueden enmarcar dentro de los siniestros obtenidos en el apartado 4.1, pero en este caso el primer y tercer clúster se enmarcarían dentro de los siniestros de severidad media con gasto elevado en pintura, mientras que el segundo y cuarto se enmarcarían dentro de los siniestros de severidad baja.

4.2.2.5. Grupo 5: golpe trasero grave

El quinto grupo encontrado con este proceso de clustering supone el más pequeño de los siete que han sido encontrados, suponiendo apenas el 4,43% de los siniestros de nuestra base de datos.

En este caso, se trata de golpes en la zona trasera del vehículo, como se puede observar con ayuda de los datos reflejados en la siguiente tabla, que muestra las piezas afectadas con mayor frecuencia en este tipo de siniestros.

	Variables significativas	Porcentaje
63303	Paragolpes trasero (zona central)	96,55%
55303	Portón trasero	85,09%
53203	Faldón trasero	57,04%
66953	Anagrama del fabricante	55,01%
94201	Faro trasero izquierdo	53,41%
94202	Faro trasero derecho	47,62%
63301	Paragolpes trasero (zona izquierda)	46,06%
63302	Paragolpes trasero (zona derecha)	40,34%
53101	Aleta trasera izquierda	23,09%
64353	Luneta trasera	22,57%
53102	Aleta trasera derecha	20,75%

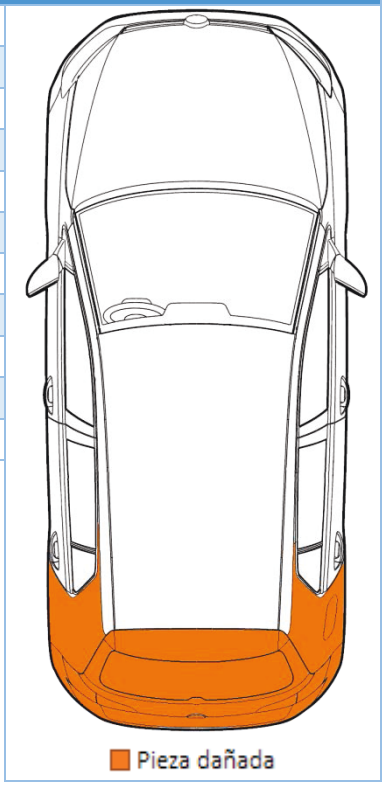


Tabla 48: zona de impacto 5: golpe trasero grave

En este grupo podemos observar como todas las piezas afectadas con frecuencia forman parte de la zona trasera del vehículo, incluyendo el paragolpes trasero en casi todos los casos, y el portón trasero en una gran parte de ellos. Además, con frecuencia resulta dañado el faldón, los faros, las aletas y la luneta de la parte trasera del vehículo.

Se puede observar como existe una homogeneidad alta en este tipo de siniestros, lo que demuestra que la mayor parte de estos siniestros tendrán varias piezas afectadas. Este tipo de siniestro puede estar ocasionado, por ejemplo, cuando el vehículo es víctima de un alcance o cuando se golpea contra una pared, columna u otro objeto a una velocidad media o alta.

Se ha realizado un análisis de ciertos parámetros estadísticos sobre los siniestros contenidos en este clúster, en concreto sobre los costes de mano de obra, pintura y sustitución de piezas de los mismos. Los resultados de este análisis se pueden observar a continuación.

	Tot_mo *Coste de mano de obra*	Tot_pint *Coste de pintura de piezas*	Tot_sust *Coste de piezas sustituidas*
Mínimo	0,00 €	0,00 €	0,00 €
Máximo	5.717,60 €	1.943,70 €	26.838,00 €
Media	409,71 €	331,58 €	567,06 €
Cuartil 1	124,80 €	205,37 €	203,16 €
Mediana	224,40 €	267,62 €	305,98 €
Cuartil 3	488,88 €	377,47 €	631,57 €
Desv. Típica	489,44 €	208,23 €	1.077,00 €

Tabla 49: zona de impacto 5: distribución de costes

Como se puede observar, el coste de mano de obra de este tipo de siniestros es por lo general medio, ya que el 50% de los siniestros tienen un coste de este tipo de entre 124,80€ y 488,88€, situándose el resto de los siniestros a ambas partes de este intervalo a partes iguales. Se puede observar que la media es bastante elevada con respecto a la mediana, lo que quiere decir, en conjunto con la elevada desviación típica (489,44€), que las observaciones más altas de esta variable serán bastante dispersos.

Por otra parte, el coste de pintura de piezas suele moverse en un rango más reducido, y sus valores también son más reducidos que en el caso anterior. Además, la desviación típica de esta variable es de 208,23€, lo cual es un valor medio, y dice que las observaciones no serán muy dispersas.

Por último, con respecto al coste de piezas sustituidas se puede afirmar que no sólo es más elevado, ya que los cuartiles y la mediana son superiores a los casos anteriores, especialmente el tercero, y además existe una desviación típica tremendamente alta (1.077,00€), lo cual significa, dada la media y el tercer cuartil, que especialmente los siniestros cuyo coste sea superior a 631,57€ serán muy dispersos.

A continuación se ha realizado un análisis de correlaciones entre las variables que se están estudiando en este clúster, el cual nos indicará cómo están relacionadas entre sí estas variables.

	Tot_mo *Coste de mano de obra*	Tot_pint *Coste de pintura de piezas*	Tot_sust *Coste de piezas sustituidas*
Tot_mo *Coste de mano de obra*	1	0,6387	0,5090
Tot_pint *Coste de pintura de piezas*	0,6387	1	0,3314
Tot_sust *Coste de piezas sustituidas*	0,5090	0,3314	1

Tabla 50: correlación entre los costes en siniestros de la zona de impacto 5

Existe una correlación moderada y positiva entre todas las variables. En este caso, es un poco superior la correlación entre el coste de mano de obra con las demás variables, pero no muy

superior. De esta manera, se espera que ante una subida en cualquiera de las variables, se espera que también se produzca una subida en el resto de las variables, pero de una forma moderada.

Se ha realizado un proceso de agrupamiento con estos datos, utilizando las variables coste de pintura, coste de mano de obra y coste de sustitución de piezas, para tratar de formar distintos grupos de severidad sobre los siniestros de este tipo. Se han formado un total de tres grupos de severidad, cuya distribución se puede observar a continuación.

Ilustración 40: distribución de la zona de impacto 5 por severidad

Estos tres grupos contienen la totalidad de los siniestros que representan un golpe trasero grave. Se han realizado una serie de análisis estadísticos sobre las variables analizadas de estos clústeres resultantes, como se puede observar en las tablas mostradas a continuación.

Clúster	Variable	Mínimo	Máximo	Media	Desv. Típica
1: Severidad media con gasto elevado en pintura	*Tot_mo*	0,00 €	854,70 €	185,09 €	122,93 €
	Tot_pint	0,00 €	1.789,80 €	306,64 €	218,05 €
	Tot_piez	0,00 €	499,00 €	105,47 €	85,98 €
2: Severidad alta	*Tot_mo*	102,30 €	3.034,50 €	756,96 €	324,34 €
	Tot_pint	0,00 €	1.843,50 €	446,56 €	193,43 €
	Tot_piez	80,17 €	26.838,00 €	882,42 €	569,87 €
3: Severidad media con gasto elevado en piezas	*Tot_mo*	23,25 €	691,31 €	199,05 €	117,31 €
	Tot_pint	0,00 €	964,60 €	236,86 €	74,06 €
	Tot_piez	178,36 €	867,43 €	359,17 €	130,17 €

Tabla 51: análisis sobre los grupos de severidad zona de impacto 5 (I)

Clúster	Variable	Q1	Mediana	Q3
1: Severidad media con gasto elevado en pintura	*Tot_mo*	98,00 €	151,94 €	236,30 €
	Tot_pint	190,01 €	237,88 €	312,72 €
	Tot_piez	32,10 €	71,37 €	191,46 €
2: Severidad alta	*Tot_mo*	518,28 €	700,63 €	945,53 €
	Tot_pint	334,41 €	399,26 €	501,69 €
	Tot_piez	647,33 €	825,56 €	1.073,70 €
3: Severidad media con gasto elevado en piezas	*Tot_mo*	108,57 €	172,24 €	266,40 €
	Tot_pint	187,12 €	230,29 €	276,54 €
	Tot_piez	266,80 €	314,77 €	406,59 €

Tabla 52: análisis sobre los grupos de severidad zona de impacto 5 (II)

En primer lugar, se puede encontrar un clúster, que contiene el 33,19% de las observaciones, donde se observan unos gastos en mano de obra y de sustitución de piezas muy reducidos, ya que el coste de mano de obra es en el 75% de los casos inferior a 236,30€ y el coste de sustitución de piezas es inferior a 191,46€ en el 75% de los casos. Adicionalmente, estas dos variables tienen una desviación típica muy reducida, lo cual indica que estas dos variables son muy estables y, por tanto, poco dispersas. Por otra parte, el gasto en pintura de piezas es el más elevado de estos tres, pero es de carácter medio, ya que el 50% de los siniestros tendrán un gasto de este tipo entre 190,01€ y 312,72€. La desviación típica de la pintura es un poco más elevada, pero no mucho más, lo que indica que esta variable en particular, y las tres en global, no serán muy dispersas en este clúster. Dada esta información, se puede afirmar que se trata de siniestros de severidad media con gasto elevado en pintura, de acuerdo a las categorías creadas en el apartado 4.1.

Por otra parte, existe un segundo clúster contiene siniestros de envergadura más grande, que contiene el 23,28% de los siniestros de tipo "golpe general". En primer lugar, se puede observar que dos de las variables, el coste de mano de obra y el de sustitución de piezas, toman mínimos distintos a cero, lo cual ya indica que estas variables tomarán valores más altos. En este caso, sólo el 25% de las observaciones tienen un coste de mano de obra inferior a 518,28€. Algo similar ocurre con los costes de sustitución de piezas, donde apenas un 25% tiene un coste de sustitución de piezas inferior a 647,33€ y un 25% superior a 1.073,70€, estando el 50% restante entre dichos dos valores. Estas dos variables además son muy dispersas dado que tienen desviaciones típicas altas. Con respecto a los costes de pintura, sin embargo, el 75% de los datos tiene un coste inferior a 501,69€, por lo que es la variable menos relevante de las tres, aunque toma un valor medio. La dispersión de esta variable es media, por lo que se espera que sea más estable. Por todo esto, se puede afirmar que se trata de un clúster que contiene siniestros de severidad alta.

Por último, existe un tercer clúster, el cual es el más grande de los tres encontrados, y que contiene el 43,52% de las observaciones de tipo golpe general. Este clúster muestra costes de pintura y de mano de obra bastante similares, donde en ambos el 75% de los siniestros se encuentran en

valores en un rango similar, [23,25€, 266,40€] y [0, 276,54€] respectivamente. El gasto en sustitución de piezas es en este caso un poco más elevado, y apenas el 25% de los siniestros tendrán gastos de este tipo inferiores a 266,80€. Además, el gasto mínimo en sustitución de piezas de este grupo es de 178,36€. Sin embargo, la desviación típica de estas tres variables es bastante baja, así que se espera que este clúster no sea muy disperso. Por todo lo anteriormente mencionado, se puede asegurar que se está hablando de siniestros de severidad media con gasto elevado en piezas, de acuerdo a la clasificación realizada en el apartado 4.1.

Adicionalmente, se ha realizado una representación de la dispersión de estos grupos en un gráfico en tres dimensiones, utilizando como ejes el coste de mano de obra, el coste de pintura y el coste de sustitución de piezas. En este caso se han limitado los tres ejes al rango [0,1.700] ya que la cantidad de valores por encima de dicho rango es muy reducida y su representación dificulta la apreciación de estos tres grupos en el gráfico.

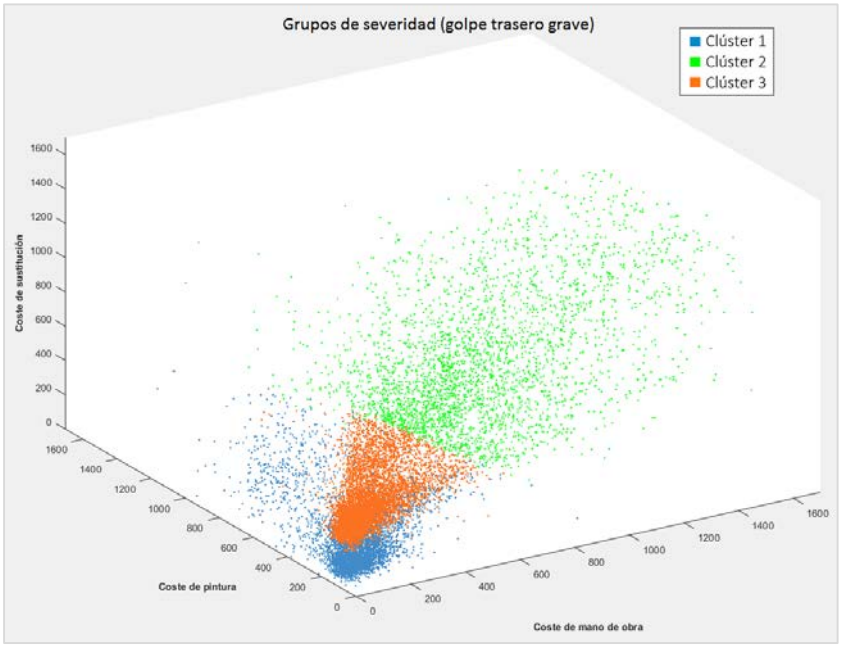


Ilustración 41: gráfico 3D de la distribución de la zona de impacto 5 por severidad

En este gráfico se pueden observar perfectamente las características de estos tres grupos, donde el primero de ellos tiene valores muy bajos y es disperso a lo largo del eje de pintura; el segundo es muy disperso en el coste de mano de obra y de sustitución, y moderadamente en los costes de pintura, y el tercero es poco disperso en las tres variables.

Estos tres grupos encajan a la perfección con los grupos de severidad obtenidos en el apartado 4.1, en concreto en los dos golpes de severidad media y el golpe de severidad alta.

4.2.2.6. Grupo 6: golpe trasero leve

Este grupo de siniestros incluye el 18,03% de los siniestros recogidos en nuestra base de datos, y recoge aquellos golpes que se han realizado en la parte trasera del vehículo pero que son de

carácter leve. A continuación se puede observar una tabla que recoge las piezas afectadas con mayor frecuencia en este tipo de golpes.

	Variables significativas	Porcentaje
63303	Paragolpes trasero (zona central)	99,99%
53101	Aleta trasera izquierda	20,37%
66953	Anagrama del fabricante	13,91%
53102	Aleta trasera derecha	12,26%
63301	Paragolpes trasero (zona izquierda)	12,04%
94201	Faro trasero izquierdo	10,15%
55303	Portón trasero	9,15%
66303	Rejilla del radiador	8,49%
66201	Molduras izquierda	7,77%
94202	Faro trasero derecho	7,39%

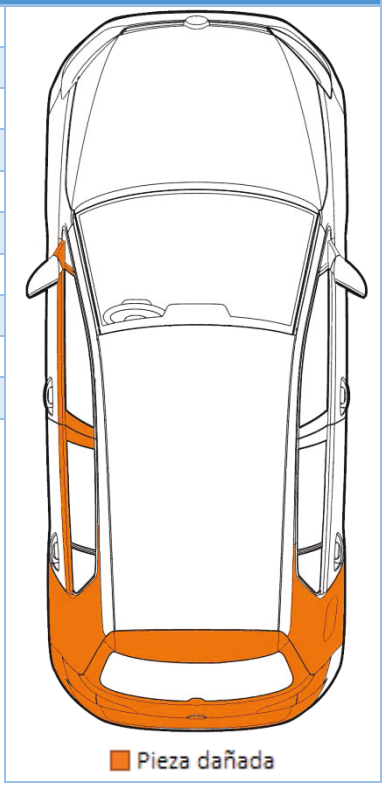


Tabla 53: zona de impacto 6: golpe trasero leve

En estos golpes resulta dañado, en casi la totalidad de los casos, el paragolpes trasero. Además, con mucha menor frecuencia, aparecen las aletas traseras, el anagrama del fabricante, los fatos traseros, el portón trasero, las molduras derechas y la rejilla del radiador. Sin embargo, se puede observar que en este grupo no existe una gran cantidad de piezas que aparezcan con una frecuencia elevada, lo que demuestra que la homogeneidad del clúster no es muy elevada. Es por ello por lo que se puede afirmar que este tipo de golpes son leves, es decir, golpes con un número reducido de piezas afectadas aunque todas se encuentran en zonas similares.

Estos golpes serán habitualmente rozaduras contra otros vehículos, paredes, muros u otros objetos, y afectarán por lo general al paragolpes trasero y, en ocasiones, a otras piezas que lo rodean.

Para continuar, se ha realizado una serie de análisis estadísticos sobre los costes que acarrea este tipo de siniestros, representados por las variables `Tot_mo`, `Tot_pint` y `Tot_sust`, para entender mejor las características de este grupo de siniestros.

	Tot_mo *Coste de mano de obra*	Tot_pint *Coste de pintura de piezas*	Tot_sust *Coste de piezas sustituidas*
Mínimo	0,00 €	0,00 €	0,00 €
Máximo	7.260,50 €	1.879,80 €	26.224,00 €
Media	115,35 €	182,78 €	129,29 €
Cuartil 1	45,50 €	107,50 €	0,00 €
Mediana	73,70 €	156,40 €	15,34 €
Cuartil 3	122,55 €	221,04 €	199,15 €
Desv. Típica	159,42 €	119,73 €	551,74 €

Tabla 54: zona de impacto 6: distribución de costes

Podemos observar que en este caso, por lo general los costes de mano de obra, pintura y sustitución de piezas son bastante bajos, dado que el tercer cuartil de todos ellos no contiene un valor muy elevado, lo cual demuestra que el 75% de los siniestros de este grupo tienen un valor igual o inferior a 122,55€ en mano de obra, 221,04€ en pintura y 199,15€ en sustitución de piezas. Adicionalmente, en la mano de obra y pintura se observa una desviación típica muy pequeña, mientras que la sustitución de piezas tiene, por lo general, una desviación típica más elevada. Además, se puede observar que al menos el 25% de estos siniestros no requieren sustitución de piezas porque el primer cuartil es 0,00€, y además también se puede afirmar que al menos el 50% de estos siniestros tienen un gasto en piezas inferior a 15,35€.

Todo esto nos demuestra que estos siniestros principalmente generarán gastos en mano de obra y pintura, y en muchas ocasiones el coste de sustitución de piezas es muy bajo o inexistente. Esto reafirma la hipótesis de que son siniestros leves, dado que las piezas se suelen reparar y pintar en lugar de ser sustituidas.

A continuación se ha realizado un análisis de correlaciones entre las variables que se están estudiando en este clúster, el cual nos indicará cómo están relacionadas entre sí estas variables.

	Tot_mo *Coste de mano de obra*	Tot_pint *Coste de pintura de piezas*	Tot_sust *Coste de piezas sustituidas*
Tot_mo *Coste de mano de obra*	1	0,5115	0,3325
Tot_pint *Coste de pintura de piezas*	0,5115	1	0,0913
Tot_sust *Coste de piezas sustituidas*	0,3325	0,0913	1

Tabla 55: correlación entre los costes en siniestros de la zona de impacto 6

Existe una correlación moderada y positiva entre el coste de mano de obra con el coste de pintura, y un poco más reducida con el coste de sustitución de piezas. De esta manera, se espera que ante una subida en el coste de mano de obra, se produzca una subida moderada en el coste de pintura y una un poco más reducida en los costes de sustitución. Lo mismo ocurre en el sentido contrario, ante una subida en el coste de pintura se producirá una subida moderada en el coste de mano de

obra, y ante una subida en los costes de sustitución se espera una subida media-baja en el coste de pintura. Por otra parte, los costes de pintura y sustitución también tienen una relación positiva, aunque muy reducida. En caso de que se produzca un incremento en una de estas dos variables, se espera que la otra también se incremente, pero en una magnitud bastante inferior.

A pesar de que los siniestros de este grupo muestran una similitud bastante alta entre sí, dada la baja desviación típica que tienen, se ha conseguido realizar una subdivisión de los mismos en 4 grupos de severidad diferentes. La distribución de los siniestros en estos cuatro grupos de severidad se puede observar a continuación.

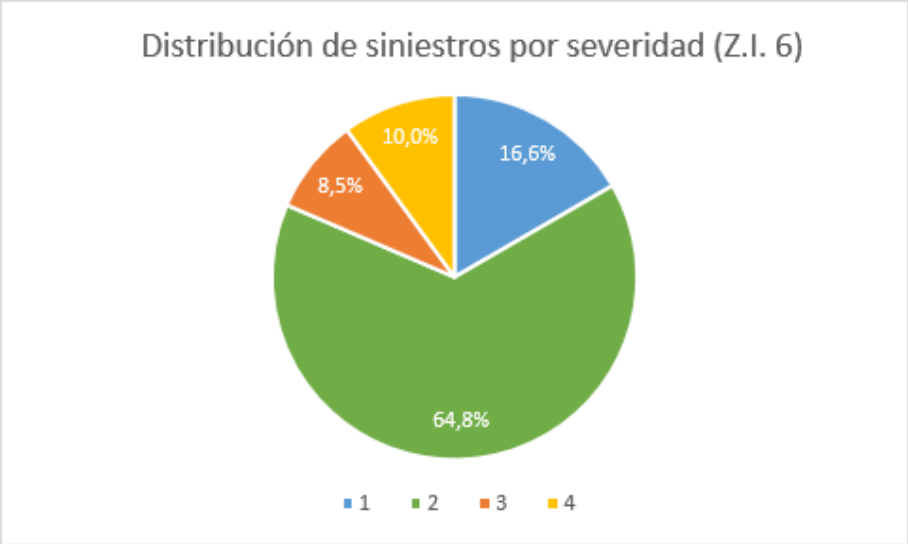


Ilustración 42: distribución de la zona de impacto 6 por severidad

Se ha generado, a partir de los clústeres obtenidos, un diagrama de dispersión en tres dimensiones que permiten apreciar a la perfección qué tipo de siniestros forma cada uno de estos grupos. Es importante tener en cuenta que se han establecido el rango para los tres ejes [0,800], ya que la cantidad de valores que hay por encima de 800,00€ en cualquiera de las tres variables es muy reducido, y hacen más difícil apreciar los grupos.

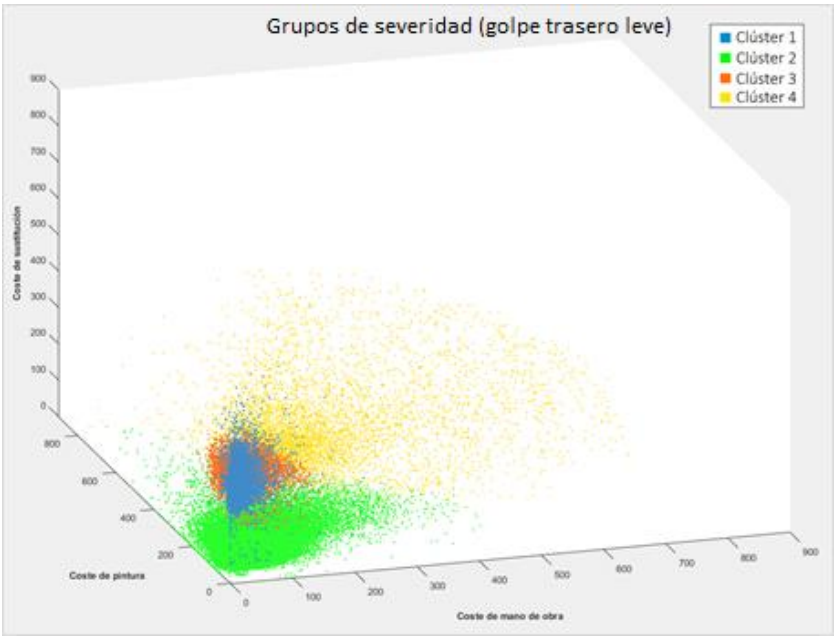


Ilustración 43: gráfico 3D de la distribución de la zona de impacto 6 por severidad

Para continuar es necesario realizar un análisis estadístico sobre cada uno de los grupos obtenidos, para averiguar qué propiedades tiene cada uno de los grupos así como comprender qué tipo de siniestros hay en su interior. Estos datos se pueden observar en las tablas a continuación.

Clúster	Variable	Mínimo	Máximo	Media	Desv. Típica
1: Severidad baja con piezas sustituidas	*Tot_mo*	0,00 €	214,20 €	44,15 €	16,45 €
	Tot_pint	0,00 €	186,01 €	99,30 €	22,53 €
	Tot_piez	51,39 €	585,37 €	217,75 €	42,70 €
2: Severidad baja con pintura y mano de obra	*Tot_mo*	0,00 €	463,58 €	88,67 €	57,31 €
	Tot_pint	0,00 €	828,47 €	174,98 €	80,06 €
	Tot_piez	0,00 €	117,22 €	9,52 €	16,75 €
3: Severidad baja con sustitución y pintura	*Tot_mo*	10,58 €	212,93 €	77,26 €	29,58 €
	Tot_pint	82,00 €	347,46 €	169,56 €	36,74 €
	Tot_piez	92,17 €	335,81 €	206,64 €	37,50 €
4: Severidad media	*Tot_mo*	0,00 €	4.162,20 €	243,32 €	145,69 €
	Tot_pint	0,00 €	842,40 €	278,97 €	120,07 €
	Tot_piez	0,00 €	26.224,00 €	268,91 €	364,52 €

Tabla 56: análisis sobre los grupos de severidad zona de impacto 6 (I)

Clúster	Variable	Q1	Mediana	Q3
1: Severidad baja con piezas sustituidas	*Tot_mo*	33,02 €	40,70 €	50,60 €
	Tot_pint	90,80 €	101,24 €	111,44 €
	Tot_piez	188,80 €	213,05 €	236,42 €
2: Severidad baja con pintura y mano de obra	*Tot_mo*	48,83 €	75,40 €	111,78 €
	Tot_pint	111,55 €	160,90 €	217,65 €
	Tot_piez	0,00 €	2,24 €	10,52 €
3: Severidad baja con sustitución y pintura	*Tot_mo*	54,90 €	72,83 €	95,70 €
	Tot_pint	143,33 €	163,76 €	192,41 €
	Tot_piez	183,51 €	205,81 €	230,48 €
4: Severidad media	*Tot_mo*	143,07 €	200,00 €	309,88 €
	Tot_pint	204,62 €	250,70 €	311,19 €
	Tot_piez	201,07 €	249,21 €	319,13 €

Tabla 57: análisis sobre los grupos de severidad zona de impacto 6 (II)

Con respecto a los grupos de severidad baja, se pueden observar tres grupos: severidad baja con piezas sustituidas, severidad baja con pintura y mano de obra y severidad baja con sustitución de piezas y pintura. Los grupos de severidad baja tienen en común que su desviación típica es muy reducida, lo cual demuestra que todos los siniestros contenidos en el mismo son muy similares entre sí. Esto se refuerza observando lo aglomerados que se encuentran los puntos de los clúster 1, 2 y 3. Estos golpes serán a menudo rozaduras leves o graves, que dependiendo de sus características y naturaleza requieren una serie de trabajos u otros.

Por ejemplo, un siniestro de severidad baja con piezas sustituidas puede suponer la rotura de molduras o piezas pequeñas que no requieran mano de obra y pintura o requieran muy pocos trabajos de ese estilo. Por otra parte, un golpe de severidad baja con pintura y mano de obra puede ser un simple raspón en una o pocas piezas de no muy alta profundidad. Por último, en un siniestro de severidad baja con sustitución y pintura, se puede predecir que habrá sido un golpe que haya roto una pieza, pero que no haya afectado a muchas alrededor.

Por otro lado, se puede observar el grupo de severidad media, donde los costes son, por lo general, de un coste más elevado, pero también tiene una desviación típica más elevada, por lo que es más disperso, y el impacto económico varía mucho más que en el caso de los siniestros de severidad baja. Este tipo de golpe tiene costes más elevados tanto en pintura, como en pintura y sustitución, lo que demuestra que se tratará de mayor gravedad, donde habrá varias piezas implicadas que requieran trabajos de los tres tipos.

Por ejemplo, este tipo de siniestro puede ser debido a un golpe frontal contra una columna, en la que resulten dañados el paragolpes delantero, un faro y la rejilla del radiador. En este caso, se podría esperar que el paragolpes sea reparado y pintado, originando costes de mano de obra y pintura, y el faro sería reemplazado, ocasionando costes de mano de obra y sustitución de piezas.

4.2.2.7. Grupo 7: golpe general

En este clúster se pueden encontrar golpes que han causado daños en piezas que se encuentran alrededor de todo el vehículo, es decir, impactos graves. Este grupo supone el 8,77% del total de los siniestros incluidos en la base de datos que se está analizando.

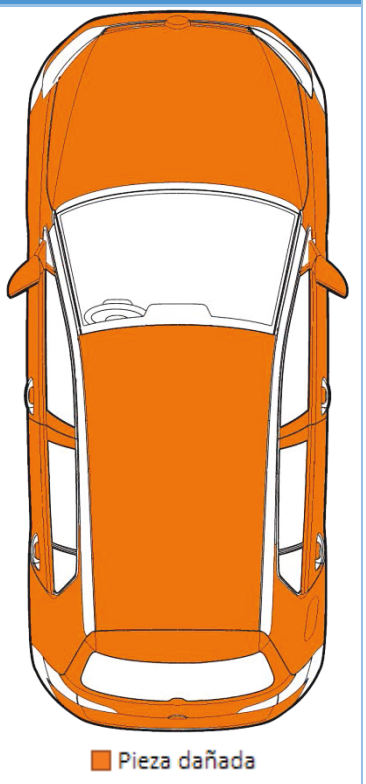
Variables significativas		Porcentaje	
53101	Aleta trasera izquierda	92,11%	
53102	Aleta trasera derecha	92,00%	
63303	Paragolpes trasero (zona central)	89,92%	
57102	Puerta delantera derecha	89,44%	
63203	Paragolpes delantero (zona central)	88,09%	
55103	Capó	87,73%	
50101	Aleta delantera izquierda	87,04%	
57101	Puerta delantera izquierda	84,39%	
55303	Portón trasero	83,03%	
50102	Aleta delantera derecha	82,98%	
66202	Molduras derecha	78,16%	
66201	Molduras izquierda	76,87%	
66303	Rejilla del radiador	73,90%	
51503	Techo	64,02%	
58102	Puerta trasera derecha	59,11%	
58101	Puerta trasera izquierda	58,14%	
66953	Anagrama del fabricante	43,50%	
66502	Retrovisor derecho	42,76%	
66501	Retrovisor izquierdo	41,94%	

Tabla 58: zona de impacto 7: impacto grave

Este tipo de siniestro incluye golpes de índole grave dado que existe una gran cantidad de piezas frecuentemente afectadas, pasando por todos los paragolpes, todas las aletas, todas las puertas, el capó, el portón trasero, las molduras, el techo, los retrovisores, etcétera. Esto demuestra que no existe una zona del vehículo que reciba la mayor parte de los impactos, sino que por lo general un impacto de este estilo afectará a numerosas piezas del vehículo distribuidas a lo largo del mismo.

Este tipo de impacto podría estar debido, por ejemplo, a la pérdida de control del vehículo, posiblemente incluyendo vueltas de campana, debido al exceso de velocidad, el estado de la calzada, un choque o roce con otro vehículo, entre otras muchas posibles causas.

Para continuar, será preciso analizar los costes generados por este tipo de siniestros, para entender de qué tipo de siniestros se trata y poder sacar algunas conclusiones.

	Tot_mo *Coste de mano de obra*	Tot_pint *Coste de pintura de piezas*	Tot_sust *Coste de piezas sustituidas*
Mínimo	0,00 €	0,00 €	0,00 €
Máximo	5.252,10 €	2.237,90 €	33.959,00 €
Media	473,16 €	936,42 €	358,37 €
Cuartil 1	242,99 €	777,86 €	37,05 €
Mediana	364,00 €	911,11 €	82,37 €
Cuartil 3	551,00 €	1.065,10 €	277,41 €
Desv. Típica	421,34 €	240,98 €	1.022,30 €

Tabla 59: zona de impacto 3: distribución de costes

A partir de los datos obtenidos en la tabla podemos observar que se trata de siniestros con, por lo general, costes medios de mano de obra, dado que la mitad de los siniestros tienen un coste de mano de obra de menos de 364,00€. La desviación típica de esta variable es bastante elevada, de forma que se puede esperar que los valores que tome no sean muy estables.

Por otra parte, los gastos de pintura son por lo general altos, ya que el primer cuartil se sitúa en 777,86€, lo que quiere decir que el 25% de los siniestros tiene un coste inferior o igual a dicha cantidad, y el 75% tiene un coste mayor. Además, este coste tiene una desviación típica baja, lo que significa que los valores se moverán por lo general en un rango pequeño.

Por último, se puede observar que los gastos de sustitución de piezas en este caso son bastante bajos, siendo el 50% de ellos inferiores o iguales a 82,37€. Aun así, si consideramos el tercer cuartil podemos ver que el 75% de los datos están por debajo de 277,41€, lo cual tampoco es un valor muy alto. Sin embargo, la desviación típica de esta variable es muy alta, situándose en 1.022,30€, lo cual indica, en conjunto con los valores analizados sobre los cuartiles, que el 25% de las observaciones restantes serán muy dispersas. De hecho, se puede observar que el valor de la media de esta variable está por encima del tercer cuartil y es 4,3 veces mayor que la media, lo que indica que la parte más alta de estos valores es muy dispersa y tomará valores muy elevados en comparación con el 75% más bajo.

A continuación se ha realizado un análisis de correlaciones entre las variables que se están estudiando en este clúster, el cual nos indicará cómo están relacionadas entre sí estas variables.

	Tot_mo *Coste de mano de obra*	Tot_pint *Coste de pintura de piezas*	Tot_sust *Coste de piezas sustituidas*
Tot_mo *Coste de mano de obra*	1	0,4472	0,6493
Tot_pint *Coste de pintura de piezas*	0,4472	1	0,1206
Tot_sust *Coste de piezas sustituidas*	0,6493	0,1206	1

Tabla 60: correlación entre los costes en siniestros de la zona de impacto 7

Existe una correlación moderada y positiva entre el coste de mano de obra con el coste de pintura, y un poco más elevada con el coste de sustitución de piezas. De esta manera, se espera que ante una subida en el coste de mano de obra, se produzca una subida moderada en el coste de pintura y una más elevada en los costes de sustitución de piezas. Lo mismo ocurre en el sentido contrario, ante una subida en el coste de pintura se producirá una subida moderada en el coste de mano de obra, y ante una subida en los costes de sustitución se espera una subida media-alta en el coste de pintura. Por otra parte, los costes de pintura y sustitución también tienen una relación positiva, aunque muy reducida. En caso de que se produzca un incremento en una de estas dos variables, se espera que la otra también se incremente, pero en una magnitud bastante inferior.

Todo esto se puede observar claramente en el gráfico en tres dimensiones que se ha dibujado sobre esta muestra. Este gráfico ha sido reducido al rango [0, 1.300] para poder apreciar mejor la distribución de la muestra, ya que existen muy pocas observaciones por fuera de este rango.

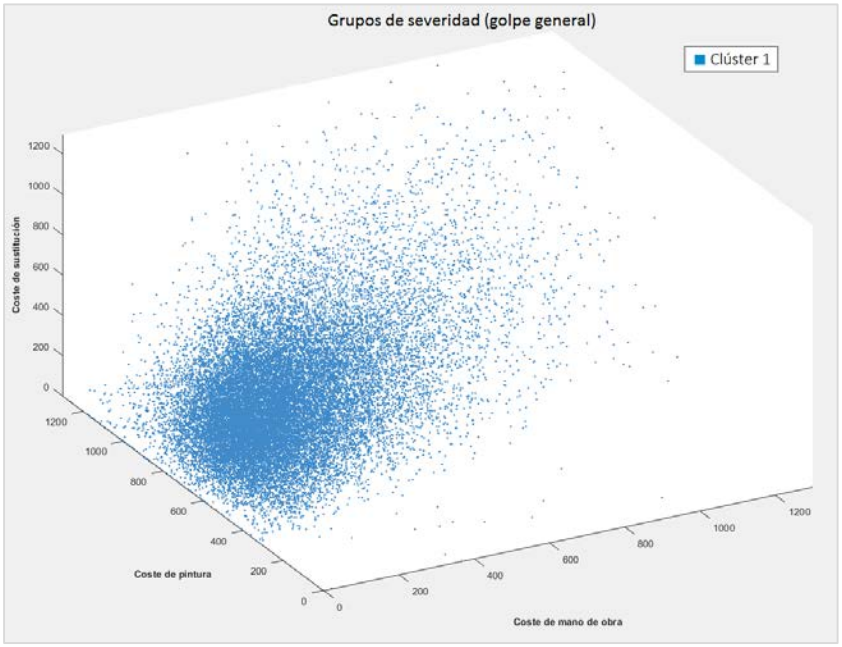


Ilustración 44: gráfico 3D de la distribución de la zona de impacto 7 por severidad

En este gráfico se ve claramente como existe una aglomeración de datos muy grande en torno a los rangos [500, 900] de coste de pintura, [0, 400] de coste de mano de obra y [0, 500] de costes de sustitución, y el resto de los partes que no se encuentran en dicho rango son muy dispersos.

Otra cosa que no se observa en las estadísticas y sí que se puede ver en este gráfico es que los costes de pintura raramente se encuentran por debajo de los 400€, por lo que se espera que un golpe de este tipo tenga un coste, como mínimo, de esta cantidad.

Por otra parte, se ha intentado realizar la subdivisión de este grupo en distintos grupos de severidad, pero esto no ha sido posible debido a que los grupos obtenidos no mostraban ningún tipo de característica común, el algoritmo de agrupamiento simplemente terminaba agrupando

los datos en clústeres de un tamaño similar los cuales, una vez representados en un gráfico en tres dimensiones, formaban cortes en línea recta de la base de datos original, lo que nos demuestra que el algoritmo ha sido incapaz de encontrar grupos con características específicas sino que simplemente hace una división, de alguna forma, a partes iguales.

Por todo esto, se puede afirmar que todos los siniestros de este tipo siguen un patrón común y son similares entre sí sin necesidad de subdividirlos en clústeres. En este caso, volviendo a la agrupación original del apartado 4.1, estaríamos hablando de que los golpes generales son habitualmente siniestros de severidad alta.

4.2.3. Conclusiones sobre las zonas de impacto

A lo largo de esta parte del estudio se han obtenido un total de siete grupos de zonas de impacto, cada uno de los cuales se dividía en entre uno y cuatro subgrupos de severidad, de manera que los siniestros quedan explicados a la perfección con este estudio.

En la siguiente página se puede observar un pequeño gráfico que describe los grupos obtenidos en forma de esquema, de forma que su comprensión sea mucho más fácil e intuitiva.

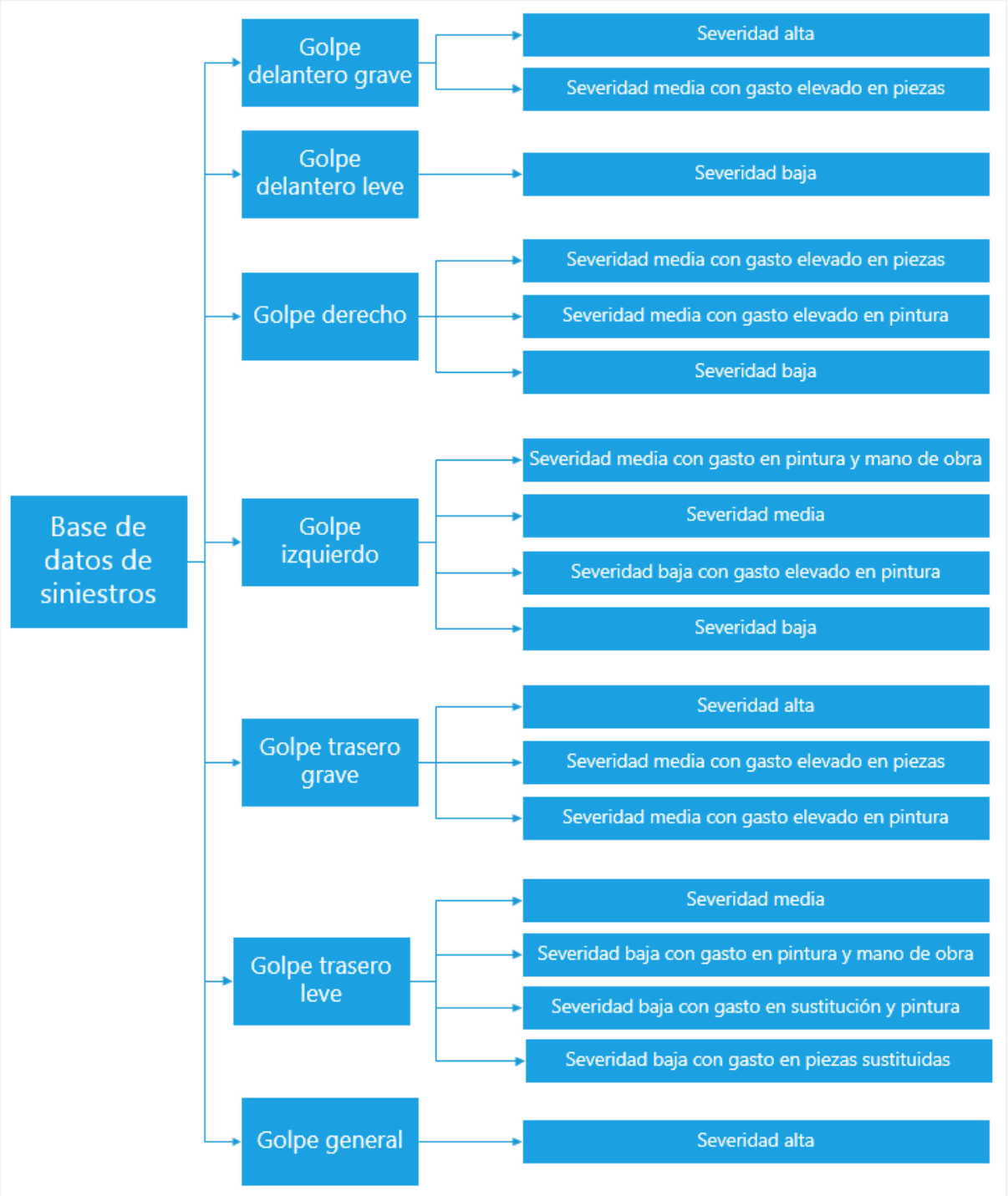


Ilustración 45: clasificación de los siniestros por zonas de impacto

4.2.4. Creación de árboles de decisión

Una vez generados los grupos de zonas de impacto, es el momento de crear árboles de clasificación, que sean capaces de categorizar nuevos siniestros con respecto a este vehículo, en cualquiera de los siete grupos de zonas de impacto de primer nivel. Para ello, al igual que en el caso anterior, se utilizará el algoritmo *Random Forests* incluido en Mahout.

Tras realizar el preprocesado necesario, se dispone de la base de datos de siniestros extendida (con las nuevas columnas de zonas de impacto que se mencionó en el apartado 4.2.2), añadiendo una columna extra que incluye una nueva variable, la cual toma valores de 1 a 7 dependiendo del grupo al que pertenezca la observación. En este caso no es fácil determinar el tipo de las variables que se van a analizar, dada la naturaleza binaria de las mismas. Por ello, se realizará el análisis de dos maneras distintas, en primer lugar considerando estas variables como textuales, y en segundo lugar como numéricas. De esta forma se podrá determinar si conviene tomarlas como numéricas, como textuales, o por el contrario si ambos conjuntos de árboles de decisión son válidos y pueden servir para reforzar la predicción.

Para estos análisis, se han considerado las siguientes variables:

- XXXXX: cada una de las 75 zonas de impacto válidas. Todas ellas aparecen en el menos el 0,1% de los siniestros, como se indicó anteriormente. Estas variables toman un valor numérico de 5 cifras de acuerdo a lo especificado en el Anexo II, y serán consideradas de tipo numérico en el apartado 4.2.4.1 y de tipo textual en el apartado 4.2.4.2.
- Clúster: número identificador del clúster de severidad de primer nivel al que pertenece el siniestro (etiqueta).

4.2.4.1. Árboles de decisión considerando las zonas de impacto como numéricas (R1)

En este proceso, el cual será identificado como *R1*, la ejecución de este proceso se realiza en un tiempo mucho más reducido que en otros casos: poco más 40 segundos. Esto es debido a que se trata de variables enteras de un conjunto predefinido, por lo que la variación entre los valores de unos y otros es muy inferior. El resultado del mismo se puede observar en la ilustración a continuación.

```
15/09/19 12:27:51 INFO common.HadoopUtil: Deleting hdfs://giaa-edge1:54310/user/hadoop/ns1-forest
15/09/19 12:27:51 INFO mapreduce.BuildForest: Build Time: 0h 0m 39s 549
15/09/19 12:27:51 INFO mapreduce.BuildForest: Forest num Nodes: 31256
15/09/19 12:27:51 INFO mapreduce.BuildForest: Forest mean num Nodes: 312
15/09/19 12:27:51 INFO mapreduce.BuildForest: Forest mean max Depth: 23
15/09/19 12:27:51 INFO mapreduce.BuildForest: Storing the forest in: ns1-forest/forest.seq
15/09/19 12:27:51 INFO driver.MahoutDriver: Program took 40539 ms (Minutes: 0.67565)
hadoop@giaa-edge1:~/TFG/Análisis/3. Arboles/2. Zonas$
```

Ilustración 46: información sobre los árboles de decisión de zonas de impacto (R1)

Se puede observar que se ha generado un bosque con 31.256 nodos, donde la longitud máxima de un árbol es de 23 nodos, y la media de nodos por árbol es de 312. Esto nos dice que existen, al igual que en el caso anterior, un total de 100 árboles, los cuales tienen una amplitud (número de hijos por nodo) muy elevada y una profundidad moderada. Esto demuestra de nuevo la complejidad del bosque generado.

Dada la complejidad de este bosque, es de gran importancia poner a prueba dichos árboles para asegurar que su fiabilidad es máxima. Una vez realizadas las pruebas, de un total de 98.651

observaciones de prueba, 86.554 han sido clasificadas correctamente, lo cual supone un 87.74% de precisión, un valor muy elevado. A continuación se puede observar la matriz de confusión, la cual nos permite desglosar estas pruebas.

	Clasificado en cl. 1	Clasificado en cl. 2	Clasificado en cl. 3	Clasificado en cl. 4	Clasificado en cl. 5	Clasificado en cl. 6	Clasificado en cl. 7
Pertenece a cl. 1	**7.199**	918	12	7	3	0	18
Pertenece a cl. 2	52	**33.302**	152	131	0	6	22
Pertenece a cl. 3	57	262	**12.833**	49	5	1	287
Pertenece a cl. 4	51	546	27	**11.663**	0	10	218
Pertenece a cl. 5	67	132	23	17	**2.975**	1.085	130
Pertenece a cl. 6	20	6.199	820	642	16	**10.086**	50
Pertenece a cl. 7	52	0	7	12	1	0	**8.496**
Tasa de acierto	*96,01%*	*80,52%*	*92,50%*	*93,15%*	*99,17%*	*90,15%*	*92,14%*

Tabla 61: matriz de confusión de los árboles de grupos de zonas de impacto (R1)

En esta tabla se puede observar, para cada clúster, el número de observaciones que fueron clasificadas en él (columnas) y el número de observaciones que pertenecen a cada clúster (filas).

Es muy fácil apreciar que el clúster número 6 (golpe trasero leve) es mal clasificado en numerosas ocasiones. El error habitualmente es asignarlo al grupo 2 (golpe delantero leve). Esto puede ser debido a que se trate de accidentes en cadena leves, por lo que se puede esperar que exista cierto error al clasificarlo en uno u otro grupo. Sin embargo, esto no suele ocurrir en sentido contrario, es decir, las observaciones del clúster 2 no suelen ser clasificados de manera incorrecta.

También se puede observar que los siniestros del grupo de golpes traseros graves (grupo 5) son clasificados por error en el grupo de golpes traseros leves (grupo 6). Esto puede ser debido a que las piezas implicadas en estos siniestros sean similares.

Adicionalmente, salta a la vista que el clúster número 2 (golpe delantero leve) recibe habitualmente por error siniestros que pertenecen a los demás grupos, especialmente del grupo 6 (golpe trasero leve) y del grupo 1 (golpe delantero grave). Esto podría ser ocasionado por accidentes en cadena, o golpes de una severidad intermedia respectivamente.

Es importante tener en cuenta que existe una tasa de acierto muy elevada en aquellos siniestros clasificados en los grupos 5 (golpe trasero grave) y 1 (golpe delantero grave), lo cual quiere decir

que aquellos siniestros clasificados en dichas categorías serán muy probablemente pertenecientes a dichos grupos.

A pesar de los errores cometidos, la tasa de acierto es muy elevada en general, realizando una categorización muy fiable. A continuación se pueden observar las variables estadísticas que se pueden extraer a partir del resultado de las pruebas realizadas.

Kappa	-6,8782
Precisión	87,74%
Fiabilidad	74,80%
σ (fiabilidad)	0,3101

Tabla 62: variables estadísticas de los árboles de grupos de zonas de impacto (R1)

Como se puede observar, la precisión de este bosque es bastante elevada, dando resultados de clasificación de siniestros muy buenos. El valor de *Kappa*, al ser negativo y elevado, indica que en este caso la precisión obtenida es muy superior a la fiabilidad esperada. La fiabilidad de este bosque es muy alta (74,80%), además de que su desviación típica es muy reducida, lo que demuestra que se trata de un modelo muy fiable y preciso.

4.2.4.2. Árboles de decisión considerando las zonas de impacto como textuales (R2)

En este proceso, el cual será identificado como *R2*, la ejecución de este proceso se realiza en un tiempo un poco más elevado que en el caso anterior, pero no mucho más: 44 segundos. El resultado del mismo se puede observar en la ilustración a continuación.

```
15/09/19 18:14:54 INFO common.HadoopUtil: Deleting hdfs://giaa-edge1:54310/user/hadoop/ns1-forest
15/09/19 18:14:54 INFO mapreduce.BuildForest: Build Time: 0h 0m 42s 798
15/09/19 18:14:54 INFO mapreduce.BuildForest: Forest num Nodes: 31060
15/09/19 18:14:54 INFO mapreduce.BuildForest: Forest mean num Nodes: 310
15/09/19 18:14:54 INFO mapreduce.BuildForest: Forest mean max Depth: 22
15/09/19 18:14:54 INFO mapreduce.BuildForest: Storing the forest in: ns1-forest/forest.seq
15/09/19 18:14:55 INFO driver.MahoutDriver: Program took 44367 ms (Minutes: 0.73945)
hadoop@giaa-edge1:~/TFG/Análisis/3. Arboles/2. Zonas$
```

Ilustración 47: información sobre los árboles de decisión de zonas de impacto (R2)

Se puede observar que se ha generado un bosque con 31.060 nodos, donde la longitud máxima de un árbol es de 22 nodos, y la media de nodos por árbol es de 310. Esto nos dice que existen, al igual que en los casos anteriores, un total de 100 árboles, los cuales tienen una amplitud (número de hijos por nodo) muy elevada y una profundidad moderada. Esto demuestra, una vez más, la complejidad del bosque generado.

Dada la complejidad de este bosque, es de gran importancia poner a prueba dichos árboles para asegurar que su fiabilidad es máxima. Una vez realizadas las pruebas, de un total de 98.651 observaciones de prueba, 85.931 han sido clasificadas correctamente, lo cual supone un 87.11% de precisión, un valor muy elevado. A continuación se puede observar la matriz de confusión, la cual nos permite desglosar estas pruebas.

	Clasificado en cl. 1	Clasificado en cl. 2	Clasificado en cl. 3	Clasificado en cl. 4	Clasificado en cl. 5	Clasificado en cl. 6	Clasificado en cl. 7
Pertenece a cl. 1	**11.161**	1	233	58	904	14	76
Pertenece a cl. 2	28	**2.333**	90	67	118	1.697	57
Pertenece a cl. 3	11	1	**8.546**	56	0	0	17
Pertenece a cl. 4	15	2	24	**7.115**	930	0	14
Pertenece a cl. 5	103	1	18	68	**33.584**	14	181
Pertenece a cl. 6	551	6	28	26	6.017	**10.337**	809
Pertenece a cl. 7	10	2	210	57	204	2	**12.855**
Tasa de acierto	*93,96%*	*99,45%*	*93,34%*	*95,54%*	*80,42%*	*85,68%*	*91,76%*

Tabla 63: matriz de confusión de los árboles de grupos de zonas de impacto (R2)

En esta tabla se puede observar, para cada clúster, el número de observaciones que fueron clasificadas en él (columnas) y el número de observaciones que pertenecen a cada clúster (filas).

Nuevamente se puede apreciar que el clúster número 6 (golpe trasero leve) es mal clasificado en numerosas ocasiones. El error más habitual es asignarlo al grupo 5 (golpe trasero grave). Esto puede ser debido a que se trate de accidentes de gravedad intermedia, donde no esté muy claro a qué grupo pertenecen. Sin embargo, esto no suele ocurrir en sentido contrario, es decir, las observaciones del clúster 5 no suelen ser clasificados de manera incorrecta con tanta frecuencia.

También se puede observar que los siniestros del grupo de golpes delanteros leves (grupo 2) son clasificados por error en el grupo de golpes traseros leves (grupo 6). Esto puede ser debido a que se trate de accidentes en cadena leves donde tanto piezas delanteras como traseras son afectadas con frecuencia.

Adicionalmente, salta a la vista que el clúster número 5 (golpe trasero grave) recibe habitualmente por error siniestros que pertenecen a los demás grupos, especialmente del grupo 6 (golpe trasero leve) y del grupo 1 (golpe delantero grave). Esto podría ser ocasionado por golpes de una severidad intermedia o accidentes en cadena respectivamente.

Es importante tener en cuenta que existe una tasa de acierto muy elevada en aquellos siniestros clasificados en el grupo 2 (golpe delantero leve), lo cual quiere decir que aquellos siniestros clasificados en dicha categoría serán muy probablemente pertenecientes al mismo.

A pesar de los errores cometidos, la tasa de acierto es muy elevada en general, realizando una categorización muy fiable. A continuación se pueden observar las variables estadísticas que se pueden extraer a partir del resultado de las pruebas realizadas.

Kappa	-6,5877
Precisión	87,11%
Fiabilidad	72,88%
σ (fiabilidad)	0,3450

Tabla 64: variables estadísticas de los árboles de grupos de zonas de impacto (R2)

Como se puede observar, la precisión de este bosque es bastante elevada, dando resultados de clasificación de siniestros muy buenos. El valor de *Kappa*, al ser negativo y elevado, indica que en este caso la precisión obtenida es muy superior a la fiabilidad esperada. La fiabilidad de este bosque es muy alta (72,88%), además de que su desviación típica es muy reducida, lo que demuestra que se trata de un modelo muy fiable y preciso.

4.2.4.3. Conclusiones sobre los árboles de decisión obtenidos

Se ha podido observar que ambos árboles generados funcionan con una precisión muy elevada, de aproximadamente el 87%. Es notorio que ambos modelos son muy buenos para clasificar los siniestros, y se podrían utilizar de manera conjunta para tomar predicciones sobre nuevos siniestros que aparezcan.

Se espera que ambos árboles den el mismo resultado a la hora de clasificar nuevas observaciones, ya que estos lo hacen con una precisión muy elevada, pero es probable que se produzcan discrepancias en caso de que uno de ellos no la categorice correctamente.

En caso de discrepancia, se podría utilizar el resultado de aquella clasificación que más precisión tenga al clasificar las observaciones en el clúster que hayan dado como resultado, de entre los dos conjuntos de árboles de clasificación creados en los apartados previos.

Para facilitar la elección de un clúster ante la discrepancia de estos resultados, se ha elaborado la siguiente tabla que permite decidir el clúster al que asignar una nueva observación. En las siguientes tablas R1 es el clúster en el que la observación es clasificada por el método R1 (apartado 4.2.4.1) y R2 es el clúster en el que una observación es clasificada por el método R2 (apartado 4.2.4.2).

R1	R2	Clúster
1	1	1
1	2	2
1	3	1
1	4	1
1	5	1
1	6	1
1	7	1
2	1	1
2	2	2
2	3	3
2	4	4
2	5	2
2	6	6
2	7	7

R1	R2	Clúster
3	1	1
3	2	2
3	3	3
3	4	4
3	5	3
3	6	3
3	7	3
4	1	1
4	2	2
4	3	3
4	4	4
4	5	4
4	6	4
4	7	4

R1	R2	Clúster
5	1	5
5	2	2
5	3	5
5	4	5
5	5	5
5	6	5
5	7	5
6	1	1
6	2	2
6	3	3
6	4	4
6	5	6
6	6	6
6	7	6

R1	R2	Clúster
7	1	1
7	2	2
7	3	3
7	4	4
7	5	7
7	6	7
7	7	7

Tabla 65: elección del clúster en caso de discrepancia entre R1 y R2

Utilizando estas tablas, por ejemplo, si el algoritmo de decisión R1 indica que debe situarse en el clúster 6, mientras que el algoritmo de decisión R2 indica que debe situarse en el clúster 7, se situará en el clúster número 6 dado que el algoritmo R1 en este caso tiene mayor precisión en aquellas observaciones categorizadas en el grupo 6 (99,45%) que las que tiene el algoritmo R2 en las observaciones categorizadas en el grupo 7 (91,76%).

5. Conclusiones

Este estudio ha permitido analizar en detalle numerosos aspectos en el área de los seguros de automóvil y de la inteligencia artificial. Se ha partido del análisis del rendimiento de tecnologías distribuidas con Mahout y Hadoop con ayuda de distintos computadores de diferentes características. Más adelante se ha profundizado en técnicas de agrupamiento y clasificación aplicadas a los seguros de automóvil. Las conclusiones obtenidas a lo largo de este estudio pueden ser tremendamente útiles para los campos de la informática y de las empresas relacionadas con los seguros, desde aseguradoras hasta empresas de peritaje.

Los resultados obtenidos en este estudio son muy satisfactorios, dado que se han conseguido los objetivos planteados, y se ha conseguido profundizar más incluso de lo que estaba planeado antes de comenzar el estudio. En resumen, este estudio ha completado con éxito las siguientes tareas:

- Análisis del rendimiento de clústeres de computadores con técnicas de minería de datos y aprendizaje automático sobre bases de datos de diferentes dimensiones.
- Creación de grupos de siniestros basados en la gravedad de los mismos, así como de subgrupos basados en los trabajos requeridos para reparar los daños producidos.
- Creación de grupos de siniestros basados en las zonas de impacto de los mismos, así como subgrupos basados en la gravedad de los siniestros de cada uno de estos grupos.
- Creación de árboles de decisión para permitir la clasificación de nuevos siniestros de manera automática por gravedad y zonas de impacto.

5.1. Desarrollo del estudio

Este estudio ha comenzado con el análisis de la viabilidad del mismo. Dicho análisis se ha realizado en dos sentidos, el primero de ellos en la viabilidad de este problema desde el punto de vista informático, y el segundo de ellos la utilidad de los resultados que se podrían obtener para las empresas del sector de los seguros. Para ello el primer paso ha sido estudiar a fondo la base de datos de peritación de siniestros de la que se disponía. Una vez estudiada dicha base de datos, se trató de comprobar qué tipo de información se podía extraer de ella así como de la utilidad de los mismos para las empresas aseguradoras, consultándolo con una experta en la materia. Adicionalmente, se dispuso de la ayuda de expertos en el campo de la Inteligencia Artificial para saber qué herramientas podían ser útiles en el estudio.

Tras la confirmación de la viabilidad del proyecto, se marcaron los objetivos del estudio, tras lo que se ha llevado a cabo un extenso periodo de documentación sobre los campos de estudio (seguros de automóvil y minería de datos) y el estado del arte. También se han estudiado y probado diversos paquetes de software para minería de datos compatibles con entornos distribuidos, así como algoritmos de agrupamiento y de clasificación.

Una vez se había determinado que se utilizarían Mahout sobre una infraestructura creada con Hadoop, se comenzó una fase muy complicada, que consistió en la puesta en marcha de las infraestructuras necesarias para llevar a cabo el estudio. Para este proyecto fueron utilizados un total de 5 computadores, prestados por el Grupo de Inteligencia Artificial Aplicada (GIAA) de la Universidad Carlos III de Madrid. Adicionalmente, se implementó una infraestructura de red local (LAN) entre todos estos computadores que permitiese una rápida comunicación entre ellos.

Se empleó una cantidad muy grande de tiempo en implementar la infraestructura Hadoop, lo cual fue tremendamente complicado dada la poca documentación disponible, probablemente debido a que se trata de tecnologías muy jóvenes. Adicionalmente, gran parte de la documentación disponible está obsoleta dada la constante evolución de estos proyectos. Se siguieron numerosos tutoriales y guías obtenidos a través de libros e Internet, con los cuales finalmente se pudo implementar a la perfección esta infraestructura, aunque esto no fue sencillo. Lamentablemente, no fue posible utilizar las últimas versiones de Mahout y Hadoop para realizar el estudio, porque la documentación era casi nula y se habían producido muchos cambios de las versiones mejor documentadas a las más actuales.

Una vez terminada la implementación de la infraestructura, comenzaron las pruebas de rendimiento, durante las cuales uno de los computadores tuvo que ser descartado porque presentaba una avería en el disco duro principal, la cual lo hacía funcionar con una latencia muy elevada, pero tardó mucho en ser detectada. Esta avería fue muy importante porque ayudó a comprender cómo afecta un nodo defectuoso a un clúster con Hadoop debido a la división equitativa de tareas que éste realiza.

Una vez completadas todas las pruebas, se comenzó el desarrollo de una pequeña herramienta en Java que permitiese ejecutar los algoritmos de agrupamiento, así como el preprocesado y el postprocesado de los datos, de una forma sencilla. Sin embargo, la API de Java de Mahout que se estaba utilizando no funcionaba adecuadamente, y se producían numerosos errores que no mostraban información adicional ni estaban documentados, por lo que se decidió descartar la utilización de esta API, dada la imposibilidad de comprender estos errores. En cambio, se decidieron utilizar llamadas a Mahout a través de la línea de comandos dentro de la herramienta en Java para suplir estos errores.

El siguiente paso fue realizar los procesos de agrupamiento con la herramienta creada, lo cual tampoco fue sencillo, ya que hubo que repetir las pruebas en numerosas ocasiones para comprobar cómo variaban los resultados así como la calidad de los mismos. El análisis de los resultados que se realizó con ayuda de Excel y Matlab, los cuales permitieron transformar los ficheros CSV resultantes, los cuales son muy extensos, en tablas con datos estadísticos. De esta forma este software ayudó enormemente a detectar rápidamente errores y cálculos imprecisos.

Aun así, se ha empleado una cantidad ingente de tiempo repitiendo pruebas y esperando resultados, que aunque no tardaban más de 15 minutos en completarse, debido a la cantidad de veces que se repitieron, esto supuso en conjunto muchísimo tiempo de espera.

Por último, se han realizado una serie de scripts en línea de comandos que realizaban el proceso de clasificación a través de los ficheros CSV que dieron como resultado los procesos de agrupamiento. Este proceso resultó bastante más sencillo que el de agrupamiento dado que en el libro de A. Gupta [8] se incluyen excelentes guías de uso e interpretación de los resultados que Mahout extrae del proceso de clasificación *Random Forest*.

A pesar de que el comienzo de este estudio ha sido muy complicado y de las numerosas dificultades que se han producido, haciendo que la duración del mismo se extendiera bastante más de lo inicialmente planificado, los resultados obtenidos han sido muy satisfactorios en todas las distintas ramas de este estudio. Asimismo, las pruebas de rendimiento han llevado a valiosas conclusiones. Por otra parte, los procesos de agrupamiento han dado resultados que muestran una gran coherencia y que son reforzados al hacer los análisis desde dos puntos de partida diferentes, y las desviaciones típicas de la mayoría de los grupos obtenidos son muy reducidas, lo que denota que los grupos se han formado de una manera muy precisa. Por otra parte, los árboles de decisión generados muestran una exactitud muy elevada al hacer pruebas sobre ellos, lo cual denota que también han sido obtenidos con éxito.

5.2. Trabajos futuros

Este estudio abre un gran abanico de posibilidades para realizar estudios relacionados y ampliar la utilidad de los resultados obtenidos, pudiendo proporcionar numerosas ventajas al sector de los seguros de automóvil.

En primer lugar, sería muy interesante realizar este mismo estudio con cada una de las diferentes marcas y modelos de vehículos de los que se disponga información, para permitir disponer de información completa y útil, y poder poner en práctica toda la teoría explicada en este trabajo. Es muy importante continuar con este estudio de esta manera para poder aplicar los resultados en la realidad de las empresas aseguradoras y de peritaje, dado que habiéndolo realizado sólo de un vehículo sólo aporta información para el mismo. Sin embargo, en este proyecto se han sentado las bases perfectas para llevar a cabo el mismo análisis en otros modelos de automóvil, lo cual requiere una cantidad grande de tiempo para ser realizado, pero las conclusiones que se pueden obtener serán claramente beneficiosas para este tipo de compañías.

Además, sería de un beneficio incalculable la creación de una base de datos común de peritaje de siniestros, donde todas las aseguradoras incluyan la información de los mismos, para así poder disponer de más cantidad de datos y obtener resultados aún más precisos.

Por otra parte, sería muy útil generar algún algoritmo que permita a las compañías aseguradoras conocer con exactitud la parte de la prima correspondiente a las reparaciones realizadas en cada marca y modelo de vehículo en función de los resultados obtenidos. Esto es apenas una parte de la prima *a priori* pero, a pesar de ser una parte pequeña del cálculo que éstos realizan, se mejorarán las estimaciones de los seguros a la hora de calcular las primas, reduciendo, entre otras cosas, el riesgo de insolvencia ajustando al máximo el coste que tiene el seguro para todos los asegurados.

Por último, sería de gran interés generar un programa informático que sea capaz de detectar automáticamente posibles fraudes en los siniestros, partiendo de las agrupaciones que han sido realizadas comprobando, mediante los árboles de clasificación generados, si el nuevo siniestro encaja en un grupo de severidad y un grupo de zonas de impacto, y mostrando una alarma en caso contrario, para que el siniestro sea investigado. Esto no sólo aumentaría la cantidad de fraudes detectados, sino que también aceleraría la detección de los mismos.

6. Planificación

En este apartado se explicará la planificación que se ha realizado para llevar a cabo este proyecto, tanto la planificación que se realizó previamente al inicio del mismo como el tiempo que realmente se ha empleado debido a los contratiempos no previstos que fueron explicados en el apartado anterior.

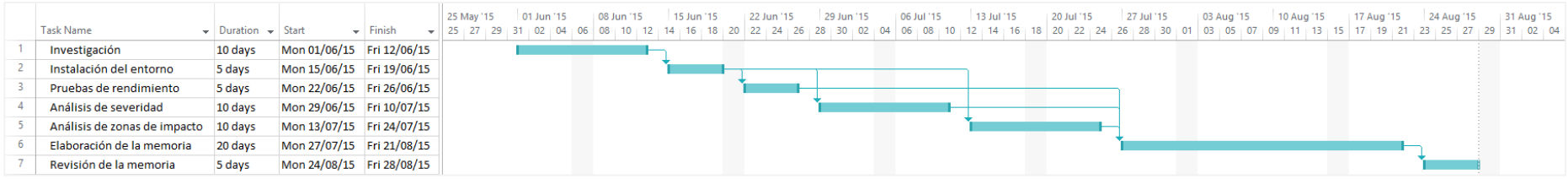
En un principio se realizó una planificación de trabajo intensivo, con plazos bastante holgados para poder afrontar cualquier contratiempo que se interpusiera en el desarrollo de este estudio. Se separó el proceso en varias fases, donde cada fase tiene asignado un número determinado de semanas, las cuales se entienden como jornada completa (40 horas):

- ***Investigación***: documentación sobre los algoritmos a utilizar, el software a emplear y la infraestructura necesaria (2 semanas).
- ***Instalación del entorno***: instalación y preparación de la infraestructura de computadores así como de Hadoop y Mahout (1 semana).
- ***Pruebas de rendimiento***: fase de medición de la velocidad y la eficiencia de los algoritmos distribuidos sobre Hadoop variando la configuración de la infraestructura (1 semana).
- ***Análisis de severidad***: análisis de los distintos grupos de severidad así como los subgrupos de zonas de impacto y los árboles de clasificación (2 semanas).
- ***Análisis de zonas de impacto***: análisis de las distintas zonas de impacto así como los subgrupos de severidad y los árboles de clasificación (2 semanas).
- ***Elaboración de la memoria***: elaboración de este texto, incluyendo todas las notas que han sido tomadas durante las anteriores fases (4 semanas).
- ***Revisión de la memoria***: repaso global del estudio para eliminar todo tipo de errores que hayan podido aparecer (1 semana).

No se espera que todas las fases sean realizadas secuencialmente, sino que se pretenden paralelizar algunas de las tareas, pero de esta forma se pueden observar las horas de trabajo efectivas planificadas en cada una de ellas. Toda esta planificación inicial hace un total de 13 semanas de trabajo, lo cual equivale a 3 meses o 520 horas de trabajo.

La planificación inicial se ha visto modificada dado que se necesitó más tiempo del inicialmente planificado debido a que los contratiempos encontrados fueron mucho más graves de lo previsto, por lo que la ejecución de las fases de instalación del entorno y análisis de severidad se han prolongado inesperadamente (1 semana cada una). Por otra parte, el proceso de análisis de zona de impacto vio mermado el tiempo empleado en el mismo dado que con la experiencia acumulada del análisis de severidad fue mucho más sencillo abordar esta fase, reduciendo el tiempo de ejecución en una semana.

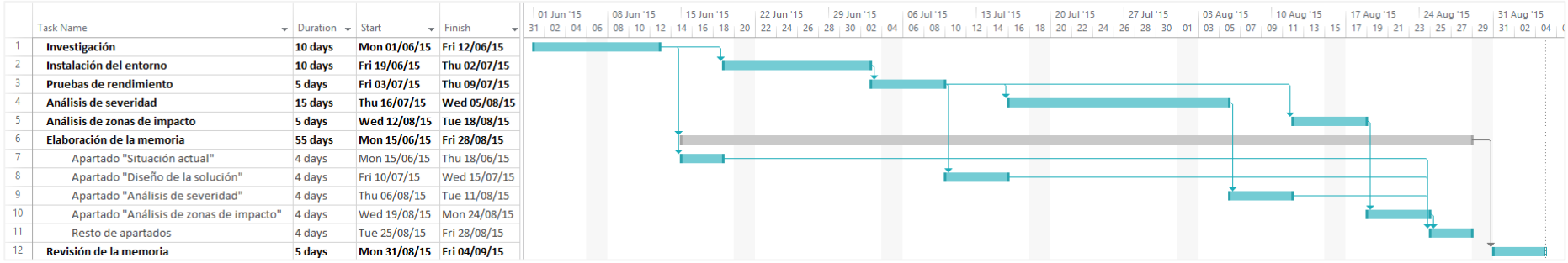
En primer lugar se puede observar la planificación inicial de este proyecto.



	Task Name	Duration	Start	Finish
1	Investigación	10 days	Mon 01/06/15	Fri 12/06/15
2	Instalación del entorno	5 days	Mon 15/06/15	Fri 19/06/15
3	Pruebas de rendimiento	5 days	Mon 22/06/15	Fri 26/06/15
4	Análisis de severidad	10 days	Mon 29/06/15	Fri 10/07/15
5	Análisis de zonas de impacto	10 days	Mon 13/07/15	Fri 24/07/15
6	Elaboración de la memoria	20 days	Mon 27/07/15	Fri 21/08/15
7	Revisión de la memoria	5 days	Mon 24/08/15	Fri 28/08/15

Ilustración 48: planificación inicial

Durante el desarrollo de las primeras fases del proyecto se decidió que sería más eficiente separar la fase de elaboración de la memoria en varias fases de un tamaño más reducido, de forma que se puedan intercalar con las demás fases del proyecto y se consiga que éste sea de mejor calidad, al escribir siempre sobre temas recientemente estudiados.



	Task Name	Duration	Start	Finish
1	**Investigación**	**10 days**	**Mon 01/06/15**	**Fri 12/06/15**
2	**Instalación del entorno**	**10 days**	**Fri 19/06/15**	**Thu 02/07/15**
3	**Pruebas de rendimiento**	**5 days**	**Fri 03/07/15**	**Thu 09/07/15**
4	**Análisis de severidad**	**15 days**	**Thu 16/07/15**	**Wed 05/08/15**
5	**Análisis de zonas de impacto**	**5 days**	**Wed 12/08/15**	**Tue 18/08/15**
6	**Elaboración de la memoria**	**55 days**	**Mon 15/06/15**	**Fri 28/08/15**
7	Apartado "Situación actual"	4 days	Mon 15/06/15	Thu 18/06/15
8	Apartado "Diseño de la solución"	4 days	Fri 10/07/15	Wed 15/07/15
9	Apartado "Análisis de severidad"	4 days	Thu 06/08/15	Tue 11/08/15
10	Apartado "Análisis de zonas de impacto"	4 days	Wed 19/08/15	Mon 24/08/15
11	Resto de apartados	4 days	Tue 25/08/15	Fri 28/08/15
12	**Revisión de la memoria**	**5 days**	**Mon 31/08/15**	**Fri 04/09/15**

Ilustración 49: planificación final

7. Presupuesto

A continuación se puede observar el presupuesto de este proyecto, incluyendo todos los costes en los que se incurre para su realización.

Para calcular el coste de infraestructura y software, se tiene en cuenta el periodo de amortización y el porcentaje de tiempo que cada una de estas herramientas es utilizada para este proyecto frente al que es utilizada en total, de manera que se calcula el coste que supone este proyecto cada uno de estos recursos.

PERSONAL					
Nombre	Coste por hora			Horas trabajadas	Coste total
Gabriel Anca Corral	22,00 €			520	11.440,00 €
				Total personal:	**11.440,00 €**
INFRAESTRUCTURA					
Nombre	Coste	Tiempo de uso	Uso proyecto	Periodo de amortización	Coste
Computador giaa-edge1	3.800,00 €	3 meses	50%	72 meses	79,17 €
Computador giaa-edge2	3.800,00 €	3 meses	75%	72 meses	118,75 €
Computador slave3	550,00 €	3 meses	100%	48 meses	34,38 €
Computador slave4	480,00 €	3 meses	100%	48 meses	30,00 €
Computador slave5	650,00 €	3 meses	50%	48 meses	20,31€
Computador para análisis	1.250,00 €	3 meses	25%	72 meses	13,02 €
Dell Networking N1500	2.295,00 €	3 meses	10%	72 meses	9,56 €
				Total infraestructura:	**305,19 €**
SOFTWARE					
Nombre	Coste	Tiempo de uso	Uso proyecto	Periodo de amortización	Coste
Microsoft Windows 8.1 Pro	128,93 €	3 meses	50%	24 meses	8,06 €
Microsoft Windows 10 Pro	230,58 €	3 meses	50%	48 meses	7,21 €
Microsoft Office 2013	222,31 €	3 meses	50%	48 meses	6,95 €
Microsoft Visio 2013	329,75 €	3 meses	100%	48 meses	20,61 €
Microsoft Project 2013	1.131,40 €	3 meses	100%	48 meses	70,71 €
Matlab R2015a	86,78 €	3 meses	100%	12 meses	21,70 €
				Total software:	**135,24 €**

TOTALES	
Concepto	Coste
Personal	11.440,00 €
Infraestructura	305,19 €
Software	135,24 €
Total sin impuestos	**11.880,23 €**
I.V.A. (21%)	2.494,89 €
Total con impuestos	**14.375,12 €**

Tabla 66: presupuesto del proyecto

8. Referencias bibliográficas

[1] Dirección General de Tráfico, «Parque de vehículos - Anuario - 2014,» 11 05 2015. [En línea]. Available: http://www.dgt.es/Galerias/seguridad-vial/estadisticas-e-indicadores/parque-vehiculos/parque_2014_anuario.xlsx. [Último acceso: 09 07 2015].

[2] Mutua Madrileña, «Mutua Madrileña,» 01 12 2014. [En línea]. Available: http://www.grupomutua.es/corporativa/InformeAnual2014/auto.jsp. [Último acceso: 21 07 2015].

[3] J. L. Pérez Torres, Fundamentos del Seguro, UMESER, 2011.

[4] «Ley sobre responsabilidad civil y seguro en la circulación de vehículos a motor,» Civitas, 2005.

[5] M. Guillén Estany, M. Ayuso Gutiérrez, C. Bolancé Losilla, L. Bermúdez Morata, I. Morillo López y I. Albarrán Lozano, El seguro de automóviles: estado actual y perspectiva de la técnica actuarial, Majadahonda: Editorial MAPFRE, 2005.

[6] J. Hernández Orallo, M. J. Ramírez Quintana y C. Ferri Ramírez, Introducción a la Minería de Datos, Madrid: Pearson Educación, 2004.

[7] E. González González, Análisis de datos aplicado a siniestros de automóviles, Colmenarejo: Universidad Carlos III de Madrid, 2015.

[8] A. Gupta, A. Hussain, R. Raman y S. Chari, Learning Apache Mahout Classification, Packt Publishing, 2015.

[9] M. G. Noll, «Running Hadoop on Ubuntu Linux (Multi-Node Cluster),» 17 07 2011. [En línea]. Available: http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-multi-node-cluster/. [Último acceso: 21 06 2015].

[10] C. Tiwary, Learning Apache Mahout, Birmingham: Packt Publishing, 2015.

[11] J. Renze, «Outlier,» 22 09 2004. [En línea]. Available: http://mathworld.wolfram.com/Outlier.html. [Último acceso: 17 07 2015].

[12] Ministerio de la Presidencia, «Boletín Oficial del Estado,» 05 11 2004. [En línea]. Available: http://www.boe.es/buscar/pdf/2004/BOE-A-2004-18911-consolidado.pdf. [Último acceso: 16 07 2015].

[13] P. Giacomelli, Apache Mahout Cookbook, Birmingham: Packt Publishing, 2013.

[14] P. Alonso González y I. Albarrán Lozano, Análisis del riesgo en seguros ne el marco de Solvencia II: Técnicas estadísticas avanzadas Monte Carlo y Bootstrapping, Majadahonda: Fundación Mapfre, 2007.

[15] Boletín Oficial del Estado, «DIRECTIVA 2009/138/CE DEL PARLAMENTO EUROPEO Y DEL CONSEJO, de 25 de noviembre de 2009, sobre el seguro de vida, el acceso a la actividad de seguro y de reaseguro y su ejercicio (Solvencia II),» 17 12 2009. [En línea]. Available: https://www.boe.es/doue/2009/335/L00001-00155.pdf. [Último acceso: 08 08 2015].

[16] Boletín Oficial del Estado, «REGLAMENTO DELEGADO (UE) 2015/35 DE LA COMISIÓN, de 10 de octubre de 2014, por el que se completa la Directiva 2009/138/CE del Parlamento Europeo y del Consejo sobre el acceso a la actividad de seguro y de reaseguro y su ejercicio (Solvencia II),» 17 01 2015. [En línea]. Available: http://www.boe.es/doue/2015/012/L00001-00797.pdf. [Último acceso: 08 08 2015].

[17] The Apache Software Foundation, «MapReduce Tutorial,» 14 11 2007. [En línea]. Available: http://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html. [Último acceso: 10 08 2015].

[18] K. Zhang, «K-means under Mahout,» 15 08 2014. [En línea]. Available: http://es.slideshare.net/kpzhangs/kmeans-under-mahout. [Último acceso: 16 06 2015].

[19] S. Pazos, M. Hurtado y C. Muravchik, «Algoritmo con compresión óptima para modelos mixtos lineales ralos,» 14 09 2011. [En línea]. Available: http://www.ing.unlp.edu.ar/investigacion/archivos/jornadas2011/ie03.pdf. [Último acceso: 03 07 2015].

[20] I. H. Witten y E. Frank, Data Mining, San Francisco: Elsevier, 2005.

[21] S. Russell y P. Norvig, Inteligencia Artificial: Un Enfoque Moderno, Madrid: Pearson Educación, 2006.

[22] J. Withanawasam, Apache Mahout essentials, Packt Publishing, 2015.

9. Anexos

9.1. Anexo I: Project's Summary in English

9.1.1. Introduction

9.1.1.1. Context

Motor vehicles are essential for the daily life of today's society as they make the transport of people and goods possible. According to the last yearbook from the Directorate General of Traffic [1], at the end of 2014 there were almost 31 million of vehicles registered in Spain. In spite of all the progress in infrastructure and security both in the vehicles and on the roads, there are still a large number of accidents which lead to personal and material damages. The risks of having these types of accidents and their uncertain cost lead to the existence of vehicle insurances.

The vehicle insurance covers the risks resulting from the car use and traffic. First, it includes coverages related to the liability of the insured for the damages caused to third persons or to their assets, although it could also include a coverage to the damages suffered by the driver, to the passengers or to the insured vehicle.

Therefore, the insured, in the event of an accident covered by the insurance policy, transfers the obligation created by the liability to the underwriter, being the one who pays for all the costs resulting from the accident. The premium paid by the insured is used to cope with the accidents, so its estimation is crucial to avoid underwriters' insolvency risks.

The intricacy of an insurance premium estimation is based on many aspects, some of them could be estimated prior to the contract signing: factors related to the vehicle, the driver or the driving; and others after the contract signing: accidents' history, fines, etc.

The most recent European Regulation elaborated by the European Commission (widely known as the Solvency Regulation II) is composed by: Directive 2009/138/EC (Solvency II) and the COMMISSION DELEGATED REGULATION (EU) 2015/35 of 10 October 2014. This Directive establishes the capital requirements depending on the risks assumed by the underwriter in order to reduce the insolvency risk. Because of the imminent entry into force of the aforementioned Directive, it is essential to improve the methods to estimate the premium charged on the insured, in order to fulfill these capital requirements, distributing adequately and fairly the premiums to all the insured people of a company.

It is expected that every insurance company manages a database comprised by accidents' surveys, so they can have an internal register with all the costs and the circumstances around each vehicle repair that they have made over time. One of the most important insurance companies in Spain, regarding vehicles, is Mutua Madrileña, that in its Annual Report-2014 [2], establishes a total of

1,304,000 accident reports within the "Vehicles" Branch, with a 12.5% market share. Even if there are only a few companies that publish this information, we can observe that the amount of information produced by the insurance companies highly increases each year. Therefore, if we take all the data from the insurance companies during the maximum time lapse possible, we would be able to draw conclusions that could be quite useful for the insurance companies. A good analysis of the information would help the insurance companies to improve the premium estimation-specially the premium estimation which is based on the vehicle's features- to identify fraud and to extract highly interesting statistical conclusions.

The data would have such a big size that it would be unmanageable for the human being and even for home or office computers. Advanced tools for data analysis are needed in order to draw conclusions in a reasonable lapse of time. Fortunately, today we live in the age called “Information Age” and so there are multiple tools that are specialized in analyzing and classifying the big amount of data generated in the world every second.

However, all these tools are in continuous development, its use has not spread yet and there is scarcely no documentation about it. Working with these types of systems is not an easy task and it requires a deep research and applied knowledge to achieve appropriate results.

9.1.1.2. Goals

An important insurance assessment company have kindly given us a database which contains detailed information about more than 390,000 accident reports that correspond with a specific vehicle model (defined by the brand, the model and the manufacturing year). Such data contain the information about each one of the pieces that had to be replaced, repaired and/or painted, as well as the repair cost, broken down into: new piece cost, labor cost and paint cost.

The present project will be based on the analysis of a database comprised by accident reports in order to group the accidents caused by each vehicle brand and model, according to two different criteria:

- Depending on the accident's severity: there will be groups based on the economic cost of the accident repair. Within these groups, there will be another division, this time seeking for prevailing impact areas in each one of the groups depending on the accident's severity.
- Depending on the impact area: there will be groups based on the impact areas that have been damaged in each accident. After the collection of those impact areas, there will be different groups depending on the accident's severity for each impact area, analyzing the cost basis for each group.

There will be as well a decision tree that will allow us to decide whether a new accident corresponds to one group or another, without executing the clustering algorithm, and also identifying fraud in an assessment.

This information has multiple uses: by the one hand, we improve the a priori estimation of the vehicle insurance premium, specially the part from the premium derived from the vehicle features, based on the groups depending on the accident's severity and the vehicle impact areas for each brand, model and manufacturing year; by the other hand, fraud identification in accident assessments will be improved with the support of a decision tree, as an accident report not corresponding with any group could imply that there was a fraud, and that should be carefully examined.

In addition, with the decision tree's support we will be able to provide feedback to the database, and then easily adding all the new accidents to such database, allowing this last one to nourish with the new information.

Given the data size that we have, for all these processes we will use clustering algorithms, which will be applied using distributed systems, seeking to know the functioning of this kind of software and how to optimize the process, so the data clustering would take the shortest time possible. As well, this would allow to make more data analyses and therefore to extract more and better conclusions. It will be also needful to evaluate the different data clustering algorithms, the decision trees' creation, and their improvements in order to use the most efficient one.

Because of the type of processes that we will carry on, this research will have as well the goal of assessing the data mining technology efficiency that are executed over the distributed computing systems. This will not only allow to know the efficiency of this type of software jointly with the aforementioned algorithms, but it will help to understand how having more or less computers of different features in the cluster affects to the efficiency, depending on the type of data that are being analyzed. Then, this will not only be applicable to the present research but also to support the decision making in other subjects.

Although the information that we have comes only from one vehicle brand and model, we will make all the efforts to make the infrastructure and software used for this research available for new databases and future accidents from other brands and models.

9.1.1.3. Regulatory Framework

As the data on which this research is based have been provided by a private entity, and that data are from individuals and companies, according to the “Personal Data Protection Organic Law 15/1999” (Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter

Personal), we have paid special attention to the deletion of all the personal data in that database, making sure that all the entries are anonymous.

Moreover, according to the aforementioned legislation and with the specific requirements by the company that provided us with the data, due to a signed confidentiality agreement, we will not include in the present research any personal information about the company.

9.1.2. Solution Design

This section describes all the decisions that have been undertaken to accomplish all the goals described in the previous section. These decisions include: which software to use, which hardware architecture is needed, in what way both of them are interconnected, the chosen clustering and classification processes, and the steps required to perform those two analyses.

The dataset on which the analyses are performed contains information about 329,013 different accidents that occurred to a certain vehicle. This information was collected during three years and contains information about accidents that happened in Spain. The dataset was exported to a CSV file that contains 719 variables about each of the registered accidents. Those variables specify the parts of the vehicle that needed to be repaired, painted and/or replaced, as well as the total costs of labor, painting and parts for each accident.

Multiple software tools are used in this project. The most important criteria used to select the different software applications were their expected performance, as we are trying to solve the problem in the most optimal way. Even though, we tried to use as many free and open source software as possible. The following software packages were chosen to perform the study:

- Apache Mahout 0.9: open-source tool that aims to build an environment for creating scalable performant machine learning applications. It includes both a Java API and a command line tool for executing multiple algorithms.
- Apache Hadoop 1.2.1: open-source software library that allows to perform distributed processing of large datasets across clusters of computers. Mahout can be easily integrated with it.
- Eclipse Mars: open-source Java IDE with Maven integration, which is required to create Mahout-based projects.
- Matlab R2015a: commercial computing environment that focuses in matrix manipulations and plotting of functions and data.
- Other software like Notepad++, Microsoft Excel, Microsoft PowerPoint and Microsoft Visio have been used for minor tasks.

- Operating Systems: Ubuntu Server 14.04 and Xubuntu Desktop 15.04 have been used on the machines performing the analyses and Microsoft Windows 8.1 and 10 have been used on the machines that execute analyses and build plots.

Once the software tools were chosen, we received a set of five computers and a network switch to be able to accomplish this study. This set consisted of two supercomputers with two 4-core processors and 64GB RAM each and three home computers with one 4-core processor and 4-8GB RAM each. These computers were set up in a local area network that allowed us to have the lowest possible latency on the tasks.

In order to perform the analyses described previously, we will use three different data mining algorithms for different purposes, which will be described onwards. The analysis process will consist on the following steps – which will be repeated for both severity analysis and impact zone analysis:

- *Data pre-processing (1).*
  First, we will ensure that the wrong observations and the outlier observations are removed from the database. Additionally, the data will be normalized to speed up the clustering processes. Finally, the input data will be converted to a suitable format for Mahout clustering algorithms.
- *Estimation of the amount of clusters.*
  The *Canopy* algorithm will be used to estimate the amount of clusters to be generated. This algorithm forms clusters based on two thresholds. The first of these thresholds determines the distance between one vector and the cluster centroid for the vector to be considered as a potential part of the cluster. The other threshold defines the distance from a vector to a cluster centroid to consider the vector as a definitive part of the cluster. However, the result of this algorithm is highly influenced by the chosen thresholds. For this reason, it is only recommended to use Canopy clustering as an estimator for the amount of clusters.
- *Generation of first-level clusters.*
  The clusters will be created using the *k-Means* algorithm, which creates a specified amount of clusters based on the repetition of two phases: assignment and update. The first of them consists on the assignment of vectors to the cluster whose centroid is closer, while the second updates the clusters centroids, so that the new centroid matches the mean of all the vectors in the cluster. After this process, one new file is created for each of the clusters containing only the observations belonging to each of them.

- *Generation of sub-clusters.*

  For each of the first-level clusters, we will find out how many sub-clusters can be generated, and create them using the exact same process as in the two previous steps – using both *Canopy* and *k-Means* algorithms described above.

- *Data pre-processing (2).*

  This step consists on preparing the data by converting it to a suitable format for Mahout classifying algorithms.

- *Generation of decision tree.*

  The decision tree will be generated using *Random Forests* algorithm, which consists on generating multiple decision trees to enhance the classification of new observations. The process consists of two steps: training (generating the decision trees) and testing (measuring the accuracy and reliability of the forest). It has been decided that each of this classification processes will use 70% of the data to train the trees and the rest for the testing phase.

9.1.3. Performance tests

It is crucial to know the performance that different clusters of computers can achieve with the selected distributed computing tools. We performed multiple tests to measure the performance and decide how many computers are needed for each type of problem and database. The following test databases were used for this purpose:

	Database 1	Database 2	Database 3	Database 4
Name	Gutenberg Ebooks	Household Power Consumption	Accidents	Bag of Words
Author	Gutenberg Project	University of California, Irvine	*Confidential*	University of California, Irvine
Content	Words	Household power consumption measurement	Vehicle accidents	List of words repetitions in documents
Observations	2,283,212	2,075,259	329,013	483,450,157
Variables	1	7	719	3
Variables type	Words	Numeric (3 decimal positions)	Numeric (binary and 2 decimal positions)	Numeric (integer)
Size	6.62MB	86.90MB	460.16MB	7.27GB
Process	WordCount	k-Means	k-Means	k-Means

Chart 67: databases used to execute the performance tests

The performance tests consist on processing these databases by different clusters of computers, which are specified in the following chart.

Configuration	Nodes
1	*1x home computer*
2	*2x home computers*
3	*3x home computers*
1s	*1x supercomputer*
2s	*2x supercomputers*
2s+2	*2x home computers + 2x supercomputers*

Chart 68: architectures used to execute the performance tests

These performance tests have been run in all the clusters specified above repeating each test 10 times in each of the clusters. The results can be seen in the following graphs, where the horizontal axis displays the clusters described before and the vertical axis displays the average execution time. The results had to be split in two different graphs to be able to appreciate the impact since the Words database execution time is highly superior to the other databases'.

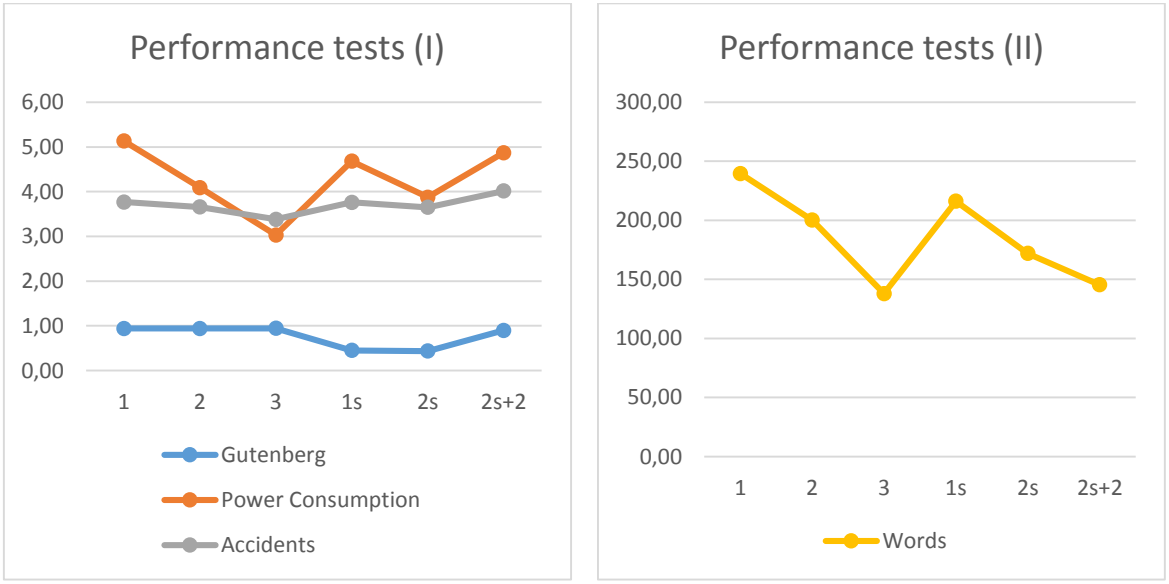


Figure 50: performance tests

According to the results from the tests, it is obvious that implementing a Hadoop architecture is only necessary for big databases. The database should have tens of millions of rows to consider using this kind of architecture, regardless of the amount of variables per row. When some extra nodes are added to the cluster, the performance is scarcely enhanced on the Gutenberg, Power Consumption and Accidents databases. However, we can appreciate a significant enhancement when some more nodes are added to the Words database, which has 483 million rows.

On the other hand it seems clear from the graphs displayed above that investing in supercomputers is probably not a good decision. A cluster formed by three home computers was more efficient than another formed by two supercomputers in all the tests executed except for the WordCount test, which should not be an example because of its reduced compute time. Given

that the total cost for the home computers cluster is €1,680 while the supercomputers cluster is worth €7,600, it is discouraged to invest in such equipment.

Finally, it is important to understand that once certain amount of nodes has been reached, the performance can either stop growing or start decreasing. For example, it has been proved that a cluster formed by three home computers is more efficient than a cluster formed by two home computers and two supercomputers. For that reason, it is highly recommended to perform some tests before deciding how many clusters to use in any study, and to add new nodes to the cluster when the database grows.

Because of these conclusions, it has been determined that three clusters will be used for this project. Each of the supercomputers will form a one-node cluster by themselves, while two home computers will form a two-node cluster. The remaining home computer will be used for monitoring purposes.

9.1.4. Severity Analysis

From an insurance company perspective, it could be very interesting to be able to categorize the accidents based on the labor, painting and part replacement costs. That would also allow them to analyze the accidents contained in each of the clusters and be able to understand the common characteristics of accidents in the same cluster.

Additionally, once those clusters are found, another important analysis would consist in the creation of different impact zone clusters for each of the severity clusters. That would make very simple to categorize and to understand the main zones that are affected on each type of impact.

After running the clustering process explained before, a total of four first-level groups have been created, which proved the existence of low severity, medium severity and high severity accidents. There are two types of medium severity accidents, depending on the highest expense (painting or part replacement). As it can be seen below, three of these groups could be subdivided in impact zone groups.

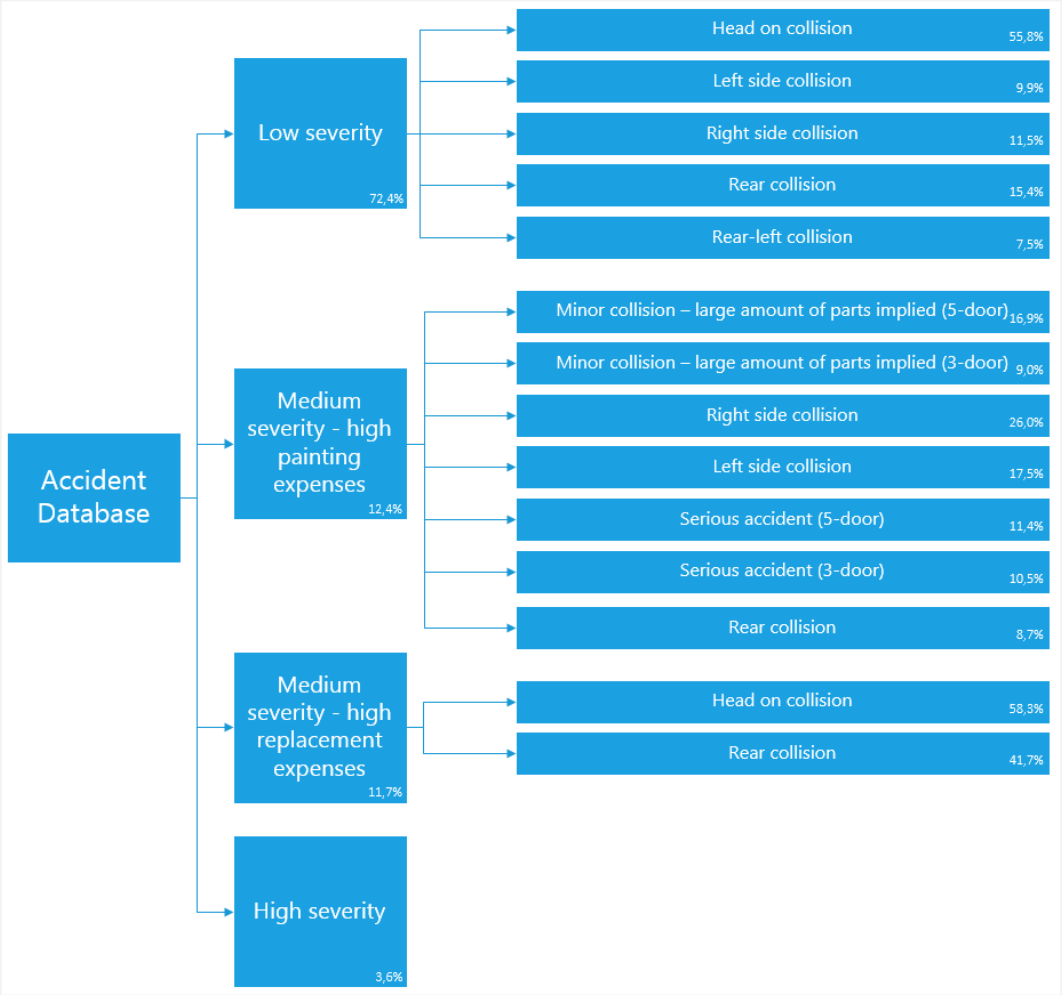


Figure 51: severity groups

Multiple statistical data had been extracted from each cluster, which proved that they show specific characteristics that define them perfectly.

Finally, 100 decision trees have been built using the *Random Forests* algorithm described above. The created forest has a total of 41,316 nodes and presents 99.71% precision and 79.42% reliability.

9.1.5. Impact Zones Analysis

This analysis intends to find out which parts of the vehicles are affected when accidents occur depending on the zone of the vehicle where there is an impact. If the most common patterns are found, insurance companies will be able to determine an estimate cost given an impact zone, as well as being able to detect potential frauds automatically when an accident is reported.

Once both *Canopy* and *k-Means* algorithms have been executed, seven different impact zone clusters have been formed. Each of these clusters can be divided in one or more severity groups, which share many characteristics with the groups formed in the severity analysis. The formed groups can be observed below.

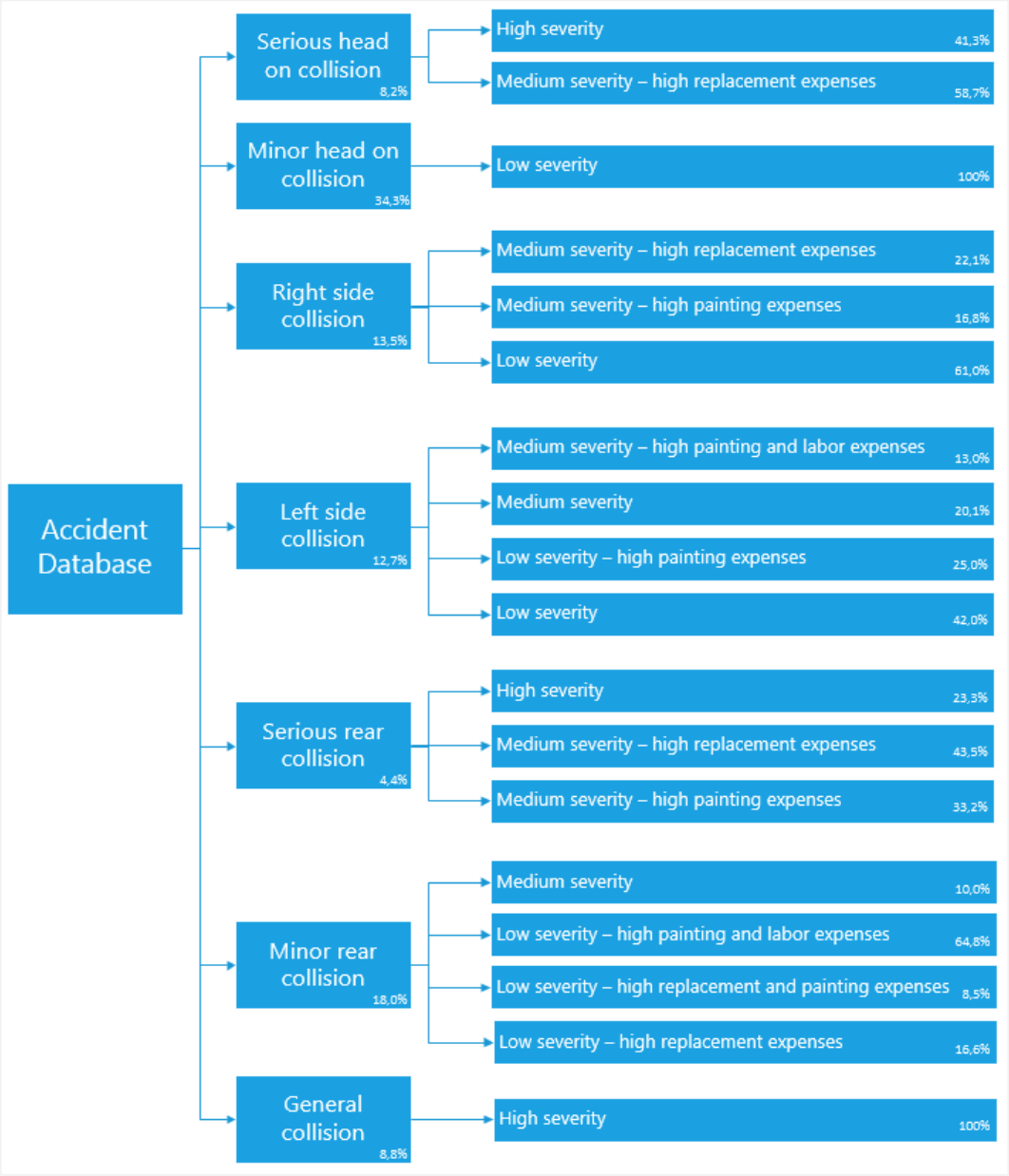


Figure 52: impact zone groups

Each of these clusters have been analyzed using multiple statistical measures that proved the existence of specific characteristics that define all of them perfectly.

Two different groups of decision trees have been created, because of their binary nature. One of the forests has been created considering the impact zones as numeric and the other considering them as discrete values. A total of 100 decision trees have been built from each point of view using the *Random Forests* algorithm described above.

The first of the forests created has a total of 31,256 nodes and presents 87.74% precision and 74.80% reliability. The second has 31,060 nodes, 87.71% precision and 72.88 reliability. After interpreting the confusion matrix we could understand that each of them is more accurate classifying in determinate clusters. Finally, it was decided that both forests would be used to

classify the new accidents into the first-level clusters. In case of discrepancy, a decision table has been created to decide which of the two results is more accurate.

9.1.6. Conclusion

The present research made possible the analysis in detail of many aspects in the vehicle insurance and in the Artificial Intelligence fields. We started from the efficiency analysis of distributed technologies with Mahout and Hadoop, with the support of various computers with different features. Then, we examined the clustering and classification techniques applied to vehicle insurances. The conclusions drawn during this research can be very useful for the IT field and also for the companies whose businesses are related to insurances, from the insurance companies to the assessment companies.

The results obtained in the present research have been satisfactory, as we have achieved the goals set out and we could deepen more than we expected before the beginning of the research. In short, this research has achieved the following goals:

- Analysis of the computer clusters' efficiency with data mining techniques and machine learning using databases of different sizes.
- Creation of groups composed by the accidents based on their severity, as well as subgroups based on all the works required to repair the damages.
- Creation of groups composed by the accidents based on the impact areas as well as on the accidents' severity from each group.
- Creation of decision trees to make possible the automatic classification of new accidents by severity and impact areas.

9.1.6.1. Research development

This research began with the analysis of its feasibility. Such analysis was made in two senses: first, about the problem's feasibility from the IT perspective, and second about the results' usefulness for the companies that work in the insurance field. For that purpose the first stage was to deeply study the database composed by the accident assessment data that we had. Once we studied such database, we tried to check what type of information could be extracted from it, as well as its usefulness for the insurance companies, having consulted with an expert in this field. Moreover, we had support from experts in the field of Artificial Intelligence in order to know what tools could be useful for the present research.

Once we had the confirmation of the project's feasibility, we set the research's goals, and after that we carried out an exhaustive documentation about the fields of study (vehicle insurances and data mining) and the state of the art. We have as well studied and tried many software suites

for compatible data mining with distributed computing environments, as clustering and classification algorithms.

Once we determined how we would use Mahout on an infrastructure created by Hadoop, we began a very difficult stage, which consisted on initiating the infrastructures needed to develop the research. For the present project we used a total of 5 computers given by the Applied Artificial Intelligence Group (GIAA) of the Universidad Carlos III (Madrid). Furthermore, we implemented a local network infrastructure between all the computers, which would allow them to have a faster communication.

It took a long time to implement the Hadoop infrastructure and it was highly complicated due to the scarce documentation that we had available, probably because these technologies are very new. In addition, a big part from the available documentation is obsolete because of the continuous development of these projects. We paid attention to some tutorials and guidelines through books and also Internet, and finally we could perfectly implement such infrastructure, even if this was not easy. Unfortunately, it was not possible to use the last Mahout and Hadoop versions in order to do the research because there was almost no documentation and there were many changes from the well documented versions to the most current one.

Once the infrastructure implementation was finished, the performance testing began, and during that stage one of the computers had to be ruled out because of a breakdown in the main hard disk, which was hard to detect, and made the computer run with a high latency. This breakdown was very important because helped us to understand how a defective node affects a cluster with Hadoop due to the equitable division of tasks made by this last one.

Once all the performance tests were completed, we began the development of a small tool in Java that would allow the clustering algorithms execution, as well as the data pre-processing and post-processing in an easy way. However, the Mahout's Java API did not function properly and there were many errors that did not show any additional information nor were recorded, so it was decided to rule out this API because it was impossible to understand such errors. Meanwhile, it was decided to use callings to Mahout through the command line inside the tool in Java in order to offset the errors.

The next stage was to do all the clustering processes with the tool that was created, which was not easy too as the tests had to be repeated many times to check how the results and their quality varied. The results analysis was made with the Excel and Matlab support, which made possible to transform the resulting CSV files, which are wide-ranging, into charts with statistical data. In this way, the software was significantly useful for promptly detecting errors and imprecise estimations. Even though, it took a long time to repeat all the tests and to wait for the results, even

if it took only 15 minutes to be completed, due to the number of times that they were repeated, this involved a long waiting time.

Finally, a series of scripts in command line were made, which carried on the classification process through the CSV files as a result from the clustering processes. This process was easier than the clustering one because in A. Gupta's book [8] there are excellent guides on how to use and interpret the results extracted by Mahout from the *Random Forest* classification process.

Despite of the difficult beginning of the present research and that there were many difficulties that prolonged it, the obtained results have been highly satisfactory for the different fields of the present research. Moreover, the efficiency tests had led to valuable conclusions. Additionally, the clustering processes have produced results that show a great coherence and which are reinforced when there is analysis made from two different starting points. The typical deviations from the majority of the obtained groups were highly reduced, which implies that the groups were precisely formed and there was some similarity between the observations within them. As well, the generated decision trees show a great accuracy after having conducted tests on them, which implies that they were also successfully obtained.

9.1.6.2. Future researches

The present research offers a wide range of possibilities to do related researches and to expand the obtained results' utility, being able to bring numerous advantages to the vehicle insurances field.

First, it would be very interesting to do this research with all the different vehicle brands and models, from which the information is available, in order to be able to dispose of complete and useful information, and as well to be able to implement all the theory explained in the present research. It is very important to continue with this research in this way in order to be able to apply the results to the reality of the insurance and assessment companies, as a research about only a vehicle just provides information for it. However, it has been shown the importance of conducting this type of researches, even if they take a long time to be developed, as the conclusions could be highly beneficial for other vehicle models' analysis.

In addition, it would be also highly advantageous the creation of a common accident assessment database, where all the insurance companies would include information about it, and then being able to dispose of more data and more accurate results.

On the other hand, it would be highly useful to create an algorithm that will allow the insurance companies to know exactly the part from the premium that corresponds to the repairs of each vehicle brand and model depending on the results that have been obtained. This is only a part from the a priori premium, but even if it is only a small part from their estimation, the insurance

estimations will be improved regarding the premiums, reducing, among others, the insolvency risk as the insurance cost for all the insured persons would be adjusted to the maximum.

Finally, it would be interesting to generate a software that would be able to automatically detect possible frauds regarding the accidents, from the clusters that have been made, and checking, through the generated classification trees, whether the new accident would correspond with the severity and impact areas clusters, and showing a warning if not, in order to examine the accident. This would not only increase the number of detected frauds, but it would also accelerate their detection.

9.2. Anexo II: Datos de la base de datos

A continuación se detallan las variables que ofrece la base de datos que se ha utilizado para el desarrollo de este estudio.

ID	Nombre	Tipo	Descripción
Secuencia	Número de secuencia	Numérico Único	Número de identificación del siniestro
Historia	Número de historia	Numérico	Número de veces que ha sido modificado el siniestro
PT_xxxxx	Mapa de bits de pintura	Binario	Para cada pieza, se almacena un valor {0, 1} que define si la pieza ha sido pintada o no. XXXXX representa el código en 5 dígitos de la pieza, que será detallado a continuación.
REP_xxxxx	Mapa de bits de reparación	Binario	Para cada pieza, se almacena un valor {0, 1} que define si la pieza ha sido reparada o no. XXXXX representa el código en 5 dígitos de la pieza, que será detallado a continuación.
SUST_xxxxx	Mapa de bits de sustitución	Binario	Para cada pieza, se almacena un valor {0, 1} que define si la pieza ha sido sustituida o no. XXXXX representa el código en 5 dígitos de la pieza, que será detallado a continuación.
Pos_int	Piezas sustituidas	Numérico	Número de piezas sustituidas
Pos_mod	Piezas reparadas	Numérico	Número de piezas reparadas
Tot_mo	Coste de mano de obra	Numérico	Coste de la mano de obra del siniestro
Tot_pint	Coste de pintura	Numérico	Coste de pintura del siniestro
Tot_sust	Coste de sustitución de piezas	Numérico	Coste de sustitución de piezas del siniestro
Tot_gen	Coste total del siniestro	Numérico	Coste total para la aseguradora del siniestro

Tabla 69: descripción de los datos de la base de datos

En las siguientes tablas se pueden observar los códigos de las piezas incluidas en la base de datos descrita con anterioridad. Para evitar prolongar innecesariamente este anexo, y facilitar la lectura del mismo, se han omitido aquellas variables sobre piezas sustituidas, pintadas o reparadas cuyo número de apariciones sea inferior al 0,1% de la muestra.

ID	Nombre	Posición
19101	Componentes del radiador	Izquierda
19102	Componentes del radiador	Derecha
19103	Componentes del radiador	Central
19201	Componentes del ventilador	Izquierda
40101	Elementos de suspensión	Izquierda
40102	Elementos de suspensión	Delantera – Derecha
40402	Eje de transmisión	Derecha
40501	Suspensión	Delantera – Izquierda
40502	Suspensión	Delantera – Derecha
42101	Elementos de suspensión	Trasera – Izquierda
42102	Elementos de suspensión	Trasera – Derecha
42103	Elementos de suspensión	Trasera – Central
44101	Neumático	Izquierda
44102	Neumático	Derecha
44201	Llanta	Izquierda
44202	Llanta	Derecha
44301	Tapacubos	Izquierda
44302	Tapacubos	Derecha
48101	Caja de dirección	Izquierda
48102	Caja de dirección	Derecha
50101	Aleta	Delantera – Izquierda
50102	Aleta	Delantera – Derecha
50203	Componentes de la coraza	Delantera – Central
51503	Techo	Central
51751	Estribo	Izquierda
51752	Estribo	Derecha
53101	Aleta	Trasera – Izquierda
53102	Aleta	Trasera – Derecha
53201	Faldón	Trasera – Izquierda
53202	Faldón	Trasera – Derecha
53203	Faldón	Trasera – Central
55101	Capó	Izquierda
55102	Capó	Derecha
55103	Capó	Central
55303	Portón	Trasera – Central
57101	Puerta	Delantera – Izquierda
57102	Puerta	Delantera – Derecha
57201	Mecanismo de cierre de puerta	Delantera – Izquierda
57202	Mecanismo de cierre de puerta	Delantera – Derecha
58101	Puerta	Trasera – Izquierda
58102	Puerta	Trasera – Derecha

Tabla 70: descripción de los códigos de las piezas (I)

ID	Nombre	Posición
63201	Paragolpes	Delantera – Izquierda
63202	Paragolpes	Delantera – Derecha
63203	Paragolpes	Delantera – Central
63301	Paragolpes	Trasera – Izquierda
63302	Paragolpes	Trasera – Derecha
63303	Paragolpes	Trasera – Central
64103	Parabrisas	Central
64201	Ventana de puerta	Delantera – Izquierda
64202	Ventana de puerta	Delantera – Derecha
64353	Luneta	Central
64401	Luna custodia	Trasera – Izquierda
64402	Luna custodia	Trasera – Derecha
64403	Luna custodia	Trasera – Central
66101	Rejilla del radiador	Izquierda
66102	Rejilla del radiador	Derecha
66103	Luna custodia	Central
66201	Molduras	Izquierda
66202	Molduras	Derecha
66303	Rejilla del radiador	Central
66501	Retrovisor	Izquierda
66502	Retrovisor	Derecha
66953	Anagrama del fabricante	Central
68101	Airbag	Izquierda
68103	Airbag	Central
68201	Cinturón de seguridad	Izquierda
68202	Cinturón de seguridad	Derecha
68203	Cinturón de seguridad	Central
70503	Tablero de instrumentos	Central
87103	Aire acondicionado	Central
92103	Limpiaparabrisas	Central
94101	Faro	Delantera – Izquierda
94102	Faro	Delantera – Derecha
94201	Faro	Trasera – Izquierda
94202	Faro	Trasera – Derecha

Tabla 71: descripción de los códigos de las piezas (II)

9.3. Anexo III: Manual de instalación de Hadoop con Mahout en múltiples nodos

La instalación de Hadoop y Mahout para crear un clúster con múltiples nodos es una tarea bastante compleja, para la que se necesitan conocimientos previos. Sin embargo, con ayuda del tutorial de M. G. Noll [9] y el libro de C. Tiwary [10] se ha realizado un detallado manual de instalación para conseguir hacer funcionar Hadoop 1.2.1 y Mahout 0.9 en cualquier clúster con nodos Linux.

En primer lugar, es importante tener en cuenta los requisitos técnicos de esta arquitectura. Será necesario disponer de uno o varios equipos Linux conectados a través de una red local o a través de internet. En este último caso, es recomendable que los equipos dispongan de direcciones IP públicas fijas para facilitar el proceso y evitar tener que cambiar el fichero *hosts* con cada cambio de IP. En cualquier caso, no se tratará la configuración de la red en este manual, se considera que se disponen de suficientes conocimientos como para configurar un router y redirigir los puertos de manera adecuada. Por otra parte, el sistema operativo recomendado es Ubuntu en sus versiones 14.04 a 15.04, ya que todo este estudio se ha desarrollado utilizando dichas versiones del sistema operativo.

Instalación y configuración del entorno en todos los nodos

En todos estos equipos, debe existir un usuario con el mismo nombre de usuario exactamente. Consideraremos que este usuario tiene como nombre `hadoop`. Si deseamos crear el usuario, debemos ejecutar los siguientes comandos:

```
$ sudo useradd hadoop -Um
$ sudo passwd hadoop
$ sudo usermod -a -G sudo hadoop
$ sudo su hadoop
```

Fragmento de código 1: creación de un nuevo usuario en Ubuntu

Con estos comandos se creará un nuevo usuario con nombre `hadoop`, un grupo con su mismo nombre y su carpeta personal en la primera línea; la segunda línea solicitará una nueva contraseña para el usuario; la tercera línea añadirá el usuario `hadoop` a la lista de *sudoers* (usuarios con privilegios de administrador) y la última permitirá al usuario identificarse en la terminal como este nuevo usuario.

Una vez tengamos en cada sistema este usuario, que será el utilizado para que cada nodo con Hadoop se comunique con los demás, debemos instalar todo el software necesario para poder hacer funcionar el entorno. Estas dependencias son: JDK de Java, Curl, SSH, RSync, OpenSSH-Server y NMap. Para instalarlos se debe ejecutar el código incluido en el fragmento de código 2.

```
$ sudo apt-get update
$ sudo apt-get install default-jdk curl ssh rsync openssh-server
nmap -y
```

Fragmento de código 2: instalación de programas necesarios para la arquitectura

Hadoop y Mahout requieren, para crear proyectos en Java, el gestor de dependencias Maven. En concreto se instalará la versión 3.2.5, que es la utilizada para este estudio, pero es probable que las versiones superiores funcionen sin problema. Para instalar Maven 3.2.5 se deben ejecutar los siguientes comandos:

```
$ cd ~
$ wget http://apache.mirrors.tds.net/maven/maven-
3/3.2.5/binaries/apache-maven-3.2.5-bin.tar.gz
$ tar zxf apache-maven-3.2.5-bin.tar.gz
$ rm apache-maven-3.2.5-bin.tar.gz
$ sudo mv apache-maven-3.2.5 /usr/local/maven
$ cd /usr/local
$ sudo chown -R $USER maven
```

Fragmento de código 3: instalación de Maven 3.2.5

El siguiente paso consiste en la instalación de Hadoop en su versión 1.2.1 (la recomendada por Mahout 0.9). De hecho, se ha comprobado durante la realización de este proyecto que las siguientes versiones de Hadoop no son compatibles con Mahout 0.9, ocasionando numerosos problemas de compatibilidad. En el fragmento de código 4 se pueden observar los comandos a ejecutar para instalar Hadoop 1.2.1.

```
$ cd ~
$ wget http://apache.rediris.es/hadoop/common/hadoop-
1.2.1/hadoop-1.2.1.tar.gz
$ tar zxf hadoop-1.2.1.tar.gz
$ rm hadoop-1.2.1.tar.gz
$ sudo mv hadoop-1.2.1 /usr/local/hadoop
$ cd /usr/local
$ sudo chown -R $USER hadoop
```

Fragmento de código 4: instalación de Hadoop 1.2.1

Para continuar, se debe descargar la versión 0.9 de Mahout de su repositorio en GitHub. Desde aquí se descargará únicamente el código fuente, que no será compilado todavía, se hará más adelante cuando todos los requisitos de Mahout estén correctamente configurados (rutas en el `PATH` de Java, Hadoop y Maven). Para descargar Mahout 0.9 se deben ejecutar los comandos indicados a continuación:

```
$ cd ~
$ wget https://github.com/apache/mahout/archive/mahout-0.9.tar.gz
$ tar zxf mahout-0.9.tar.gz
$ rm mahout-0.9.tar.gz
$ sudo mv mahout-mahout-0.9 /usr/local/mahout
$ cd /usr/local
$ sudo chown -R $USER mahout
```

Fragmento de código 5: descarga de Mahout 0.9

Como habíamos mencionado anteriormente, será necesario añadir a la variable de sistema `PATH` las rutas de Hadoop, Java, Maven y Mahout. De esta forma se podrán ejecutar estos cuatro programas sin necesidad de especificar la ruta completa de los mismos.

```
$ cd ~
$ echo "export HADOOP_HOME=/usr/local/hadoop" >> .bashrc
$ echo "export HADOOP_CONF_DIR=\$HADOOP_HOME/conf" >> .bashrc
$ echo "export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-amd64" >>
.bashrc
$ echo "export MVN_HOME=/usr/local/maven" >> .bashrc
$ echo "export MAHOUT_HOME=/usr/local/mahout" >> .bashrc
$ echo "export
PATH=\$PATH:\$JAVA_HOME/bin:\$HADOOP_HOME/bin:\$MVN_HOME/bin:\$MA
HOUT_HOME/bin" >> .bashrc
$ bash
```

Fragmento de código 6: adición de rutas a la variable PATH

El último comando del fragmento de código anterior cargará de nuevo el intérprete de comandos `bash`, permitiendo que se carguen de nuevo las variables de sistema. Es imprescindible que se ejecute este comando o se reinicie la terminal antes de proceder con el siguiente paso.

El siguiente paso consiste en la compilación tanto de Mahout como del código de ejemplo que incluye, ya que este último contiene algunos ficheros que serán de utilidad para Hadoop y las distintas ejecuciones de la API de Java de Mahout. Se debe ejecutar línea a línea todo lo incluido en las siguientes líneas de código, ya que Maven (`mvn`) escribe la salida a través de la salida para errores de la consola, y por lo general todos los intérpretes de comandos se detendrán después de la ejecución de ambas líneas por seguridad al haber recibido texto a través de la salida de errores. Este proceso será muy lento ya que Maven descargará todas las dependencias de Mahout previamente a la compilación.

```
$ cd /usr/local/mahout
$ mvn install
$ cd examples
$ mvn install
```

Fragmento de código 7: compilación de Mahout

Para continuar, es necesario que cada nodo disponga de una clave SSH, que será utilizada más adelante para registrar el equipo en los demás nodos de manera que se puedan conectar entre ellos sin necesidad de especificar la contraseña. Esto es lo que permitirá a Hadoop realizar de forma distribuida los procesos MapReduce.

```
$ ssh-keygen -t rsa -P ""
$ cat $HOME/.ssh/id_rsa.pub >> $HOME/.ssh/authorized_keys
```

Fragmento de código 8: creación de claves SSH

Es conocido que la versión 1.2.1 de Hadoop tiene ciertos problemas de compatibilidad con las redes IPv6. En este caso, como no es necesario utilizar IPv6 para ningún otro

software de la máquina, lo desactivaremos en el sistema operativo, pero también puede ser desactivado a través de los parámetros de configuración de Hadoop.

```
$ sudo echo "net.ipv6.conf.all.disable_ipv6 = 1" >>
/etc/sysctl.conf
$ echo "net.ipv6.conf.default.disable_ipv6 = 1" >>
/etc/sysctl.conf
$ echo "net.ipv6.conf.lo.disable_ipv6 = 1" >> /etc/sysctl.conf
```

Fragmento de código 9: desactivación de IPv6

Los comandos detallados en el fragmento de código 10, añaden ciertos parámetros y funciones que son necesarios para la ejecución de Hadoop en sistemas distribuidos.

```
$ echo "" >> $HOME/.bashrc
$ echo "unalias fs &> /dev/null" >> $HOME/.bashrc
$ echo "alias fd=\"hadoop fs\"" >> $HOME/.bashrc
$ echo "unalias hls &> /dev/null" >> $HOME/.bashrc
$ echo "alias hls=\"fs -ls\"" >> $HOME/.bashrc
$ echo "" >> $HOME/.bashrc
$ echo "lzohead () {" >> $HOME/.bashrc
$ echo "      hadoop fs -cat $1 | lzop -dc | head -1000 | less"
>> $HOME/.bashrc
$ echo "}" >> $HOME/.bashrc
$ echo "" >> $HOME/.bashrc
$ bash
```

Fragmento de código 10: configuración de bashrc

Para continuar, se debe especificar en el fichero hosts del sistema la IP correspondiente a cada uno de los equipos de este clúster. En este caso tendremos un computador maestro (`master`) y dos esclavos (`slave1` y `slave2`). Los comandos a ejecutar son los siguientes:

```
$ sudo echo "192.168.1.2 master" >> /etc/hosts
$ sudo echo "192.168.1.3 slave1" >> /etc/hosts
$ sudo echo "192.168.1.4 slave2" >> /etc/hosts
```

Fragmento de código 11: actualización del fichero hosts

El siguiente paso de este manual consiste en configurar todas las variables del entorno de Hadoop así como modificar los ficheros de configuración para asegurarnos de que el funcionamiento del mismo es correcto. En primer lugar configuramos las variables de entorno: la ruta del JDK de Java y el tamaño máximo de memoria disponible para las instancias de Hadoop en Java.

```
$ sudo echo "export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-amd64"
>> /usr/local/hadoop/conf/hadoop-env.sh
$ sudo echo "export HADOOP_HEAPSIZE=4096" >>
/usr/local/hadoop/conf/hadoop-env.sh
```

Fragmento de código 12: variables de entorno de Hadoop

Debemos modificar el fichero `/usr/local/hadoop/conf/mapred-site.xml` estableciendo el nombre y el puerto del computador donde se debe reportar el progreso del proceso

MapReduce. El computador encargado de recibir esta información es `master`, ya que es el que coordina el proceso completamente.

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>
 <property>
       <name>mapred.job.tracker</name>
       <value>master:54311</value>
       <description>The host and port that the MapReduce job
tracker runs at. If "local", then jobs are run in-process as a
single map and reduce task.</description>
 </property>
</configuration>
```

Fragmento de código 13: fichero de Hadoop mapred-site.xml

El siguiente fichero a modificar es `/usr/local/hadoop/conf/core-site.xml`, donde se establece la ruta temporal de Hadoop (la hemos creado con anterioridad) así como el nombre y el puerto del computador donde se encuentra el sistema de ficheros distribuido. El sistema de ficheros distribuido, en este caso, también se encontrará en el computador `master`.

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>
 <property>
       <name>hadoop.tmp.dir</name>
       <value>/var/hadoop/tmp</value>
       <description>A base for other temporary
directories.</description>
 </property>
 <property>
       <name>fs.default.name</name>
       <value>hdfs://master:54310</value>
       <description>The name of the default file system. A URI
whose scheme and authority determine the FileSystem
implementation. The uri's scheme determines the config property
(fs.SCHEME.impl) naming the FileSystem implementation class. The
uri's authority is used to determine the host, port, etc. for a
filesystem.</description>
 </property>
</configuration>
```

Fragmento de código 14: fichero de Hadoop core-site.xml

En este momento se debe crear el directorio temporal que acaba de ser especificado en el fichero `core-site.xml`, para que Hadoop sea capaz de crear y modificar sus ficheros temporales. El propietario y el grupo de dicho directorio debe ser el usuario que utilizará Hadoop para realizar el proceso. Se deben ejecutar los comandos especificados en el fragmento de código 15.

```
$ sudo mkdir -p /var/hadoop/tmp
$ sudo chown -R hadoop:hadoop /var/hadoop
```

Fragmento de código 15: creación del directorio temporal de Hadoop

Para continuar, se debe especificar en el fichero de configuración del sistema de ficheros distribuido de Hadoop (`/usr/local/hadoop/conf/hdfs-site.xml`) la cantidad de nodos en la que se deben replicar los ficheros del sistema de ficheros distribuido. En el caso de nuestro ejemplo este número es 3 (`master`, `slave1` y `slave2`).

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>
 <property>
       <name>dfs.replication</name>
       <value>3</value>
       <description>Default block replication. The actual number
of replications can be specified when the file is created. The
default is used if replication is not specified in create
time.</description>
 </property>
</configuration>
```

Fragmento de código 16: fichero de Hadoop hdfs-site.xml

Para concluir la instalación de Hadoop se debe establecer qué equipo es el maestro. De esta forma, todos los computadores sabrán quién es el que dirige el proceso. Para esto debemos escribir master en el fichero `/usr/local/hadoop/conf/masters`. El fichero de esclavos `/usr/local/hadoop/conf/slaves` contendrá en primer lugar el nombre de la propia máquina. En el caso del equipo maestro, habrá que añadir más líneas a dicho fichero, pero esto será detallado con posterioridad.

```
$ sudo echo "master" > /usr/local/hadoop/conf/masters
$ cat /etc/hostname > /usr/local/hadoop/conf/slaves
```

Fragmento de código 17: configuración del fichero masters y slaves

Habitualmente Ubuntu tiene el firewall desactivado por defecto, pero en cualquier caso conviene asegurarse de que los puertos necesarios para la ejecución de este programa están abiertos (54310 y 54311). Esto puede ser llevado a cabo con los siguientes comandos:

```
$ sudo iptables -I INPUT -p tcp --dport 54310 -j ACCEPT
$ sudo iptables -I INPUT -p tcp --dport 54311 -j ACCEPT
$ sudo iptables -F
```

Fragmento de código 18: apertura de puertos de Hadoop

Para continuar debemos configurar el acceso sin contraseña en todos los equipos. Esto nos permitirá conectarnos entre los distintos nodos sin requerir confirmación ni contraseña. Este proceso se realiza ejecutando los comandos que se pueden observar en el fragmento de código 19.

```
$ ssh-copy-id -i $HOME/.ssh/id_rsa.pub hadoop@master
$ ssh-copy-id -i $HOME/.ssh/id_rsa.pub hadoop@slave1
$ ssh-copy-id -i $HOME/.ssh/id_rsa.pub hadoop@slave2
```

Fragmento de código 19: divulgación de claves SSH

Por último, es preciso reiniciar todas las máquinas para asegurarnos de que todo el proceso se ha completado correctamente.

Configuración del nodo maestro

Para terminar, sólo es necesario realizar una serie de ajustes en el nodo maestro para, en primer lugar, establecer cuáles son sus esclavos y, en segundo lugar, inicializar el sistema de ficheros distribuido.

En primer lugar, estableceremos los esclavos en el fichero de esclavos de Hadoop (`/usr/local/hadoop/conf/slaves`):

```
$ echo "slave1" >> /usr/local/hadoop/conf/slaves
$ echo "slave2" >> /usr/local/hadoop/conf/slaves
```

Fragmento de código 20: configuración de esclavos en el nodo maestro

Por último, formatearemos el sistema de ficheros distribuido a través del siguiente comando:

```
$ hadoop namenode -format
```

Fragmento de código 21: formateo del sistema de ficheros distribuido

Puesta en marcha y pruebas de Hadoop

Para poner en marcha Hadoop con la configuración recientemente terminada, debemos iniciar en primer lugar el sistema de ficheros distribuido, y después el proceso MapReduce, como se puede observar a continuación:

```
$ $HADOOP_HOME/bin/start-dfs.sh
$ $HADOOP_HOME/bin/start-mapred.sh
```

Fragmento de código 22: puesta en marcha de Hadoop

Para comprobar si Hadoop ha arrancado correctamente se debe ejecutar el siguiente comando en todas las máquinas:

```
$ jps
```

Fragmento de código 23: listado de procesos Java en ejecución

Este comando, en la máquina maestra debe mostrar, al menos, los procesos TaskTracker, SecondaryNameNode, JobTracker, DataNode y NameNode. En las máquinas esclavas debe mostrar, al menos, TaskTracker y DataNode, pero no debe mostrar los que fueron especificados sobre el maestro (SecondaryNameNode, JobTracker o NameNode).

Para verificar que el sistema de ficheros distribuido funciona podemos crear un fichero y verificar que es accesible desde todos los computadores, utilizando el código que se puede observar a continuación:

```
$ cd ~
$ echo "esto es una prueba" > prueba_hadoop.txt
$ hadoop dfs -copyFromLocal prueba_hadoop.txt prueba_hadoop.txt
$ hadoop dfs –ls
$ hadoop dfs –cat prueba_hadoop.txt
```

Fragmento de código 24: prueba del sistema de ficheros distribuido

Tras la ejecución de estos comandos se debería mostrar en el terminal en primer lugar un fichero con ruta `/user/hadoop/prueba_hadoop.txt` y en el segundo el texto "`esto es una prueba`", sin comillas.

Para realizar pruebas de que MapReduce está funcionando correctamente, se recomienda visitar el tutorial de M. G. Noll [9], donde en el apartado "*Running a MapReduce job*" se explica detalladamente como llevar a cabo un trabajo MapReduce.

En el caso de Mahout, se dispone de varios ejemplos en el propio directorio de Mahout (`/usr/local/mahout/examples/bin/`), con los cuales se puede verificar su correcto funcionamiento.

Tras este proceso, si se desea desarrollar aplicaciones utilizando la API de Mahout, es recomendable la instalación del entorno de desarrollo *Eclipse IDE for Java EE Developers* en una de las máquinas, el cual se puede descargar de su página web (https://eclipse.org). Una vez instalado, se debe instalar el plugin *m2e*, que permitirá crear y compilar proyectos a través del gestor de dependencias Maven. Para una información más detallada se recomienda leer el apartado "*Configuring Eclipse with the Maven plugin and Mahout*" en el libro de C. Tiwary [10], donde viene explicado en detalle. Nosotros nos hemos centrado aquí únicamente en la instalación del entorno.